Multiscale Modelling of Protein Aggregation, Peptide/Membrane Interactions and Hydrophobic Core Stabilization

Dissertation
zur
Erlangung der naturwissenschaftlichen Doktorwürde
(Dr. sc. nat.)

vorgelegt der
Mathematisch-naturwissenschaftlichen Fakultät
der
Universität Zürich

von
Riccardo Pellarin
aus
Italien

Promotionskomitee

Prof. Dr. Amedeo Caflisch
Prof. Dr. Raimund Dutzler

Zürich 2007

A Nathalie

1 Summary

Biochemistry and molecular biology are directly related to everyday life: the knowledge of the biological mechanisms at the molecular level has a strong impact on health, environment and technological development.

In the present thesis, the contributions of computational methods to molecular biology were investigated. Thanks to the intrinsic flexibility of numerical simulations, a variety of complex molecular phenomena have been addressed. One of the major issues of molecular dynamics simulations is how to efficiently sample the conformational space of complex systems with a sufficiently accurate force field. The relationship between accuracy and efficiency must be established by the requirements of the investigated system. In this work three subjects were treated: amyloid aggregation, stabilization of protein folded state, and membrane-protein interactions. All the projects exposed in this thesis were approached using different computational methods with different levels of simplification and accuracy.

Amyloid protein aggregation is related to many degenerative diseases, such as Alzheimer's disease, Parkinson and type II Diabetes. Therefore the understanding of this process at the molecular level can help to devise strategies against the toxicity induced by amyloid deposits. The prediction of the aggregation rate of amyloid peptides was addressed using a phenomenological formula that gave insights into the dependence of amyloidogenesis on the polypeptide sequence. Thereafter a coarse-grained model of amyloid peptides was developed. Here, the simplification introduced allowed the exploration of phenomena such as oligomer formation, fibril nucleation and fibril elongation, which are not accessible by accurate force fields. Furthermore, the change of a single energetic parameter, which determines the relative population of the amyloid-prone and amyloid-protected states of the monomer, allowed to reproduce a vast phenomenology, which is useful to shed light into the kinetics of aggregation.

Protein folding is the phenomenon that lies behind the functionality of molecular components of the cell. Predicting which sequence modifications results in a more stable fold, and possibly more functional protein, has an enormous influence on molecular-

based technologies. Here, a collaboration with experimentalists has led to the discovery of the mutations that stabilize the folded state of an Armadillo repeat protein, which previously displayed molten globule-like features. The computational contribution helped to restrict the sequence space, and to select the most promising mutants.

Membrane-protein interactions are the basis for cell signalling and cellular transport. These mechanisms, although involved in a number of vital processes, are still poorly understood at the molecular level. In this work, the investigation of peptides that interact with both micelles and membrane by means of molecular dynamics simulations has provided new hints on the underlying molecular mechanisms. In the first system investigated the spontaneous folding of melittin on the lipid micelle surface was reproduced, together with equilibrium properties of the micelle-melittin complex. In the second work the lipid modified C-terminal heptapeptide of the human N-ras protein embedded into a membrane bilayer was studied. The simulation results validated a previous structural model based on spectroscopic data, and propose a mechanism for peptide insertion into the membrane

2 Zusammenfassung

Biochemie und Molekularbiologie stehen im direkten Zusammenhang mit dem täglichen Leben: das Wissen von biologischen Mechanismen hat eine grosse Bedeutung für die Gesundheit, die Umwelt und die technologische Entwicklung.

Die vorliegende Doktorarbeit befasst sich mit computergestützten Beiträgen zur Molekularbiologie. Eine Vielzahl von komplexen molekularen Phänomenen wurde dank der inhärenten Flexibilität numerischer Simulationen untersucht. Ein zentrales Problem von Simulationen der Moleküldynamik ist die Frage, wie der Konformationsraum eines komplexen Systems mit einem genügend genauen Kraftfeld abgetastet werden soll. Das Verhältnis zwischen Genauigkeit und Effizienz ergibt sich hierbei aus den Anforderungen des zu untersuchenden Systems. In dieser Arbeit wurden drei Themen untersucht: Amyloid Aggregation, Stabilisierung des gefalteten Zustands eines Proteins und Membran-Protein Interaktionen. Alle Projekte dieser Arbeit wurden mit verschiedenen computergestützten Methoden mit unterschiedlichem Grad an Vereinfachung und Genauigkeit behandelt.

Die Aggregation von Amyloid Proteinen wird mit vielen degenerativen Krankheiten wie Alzheimer, Parkinson und Typ II Diabetes, in Verbindung gebracht. Aus diesem Grund kann das Verständnis der Vorgänge auf molekularer Ebene dazu genutzt werden, Strategien gegen die Giftigkeit der Amyloidablagerung zu entwickeln. Die Vorhersage der Aggregationsgeschwindigkeit von Amyloidproteinen wurde mit Hilfe einer phänomenologischen Formel untersucht, welche Aufschluss gab über die Abhängigkeit der Bildung von amyloiden Strukturen aufgrund der Polypeptidesequenz. Danach wurde ein grobkörniges Modell von Amyloidpeptiden entwickelt. Die eingeführte Vereinfachung erlaubte die Erforschung von Phänomenen wie Oligomerbildung, Keimbildung und Verlängerung von Fibrillen, die mit genauen Kraftfeldern nicht zugänglich sind. Des Weiteren liess die Änderung eines einzigen energetischen Parameters, welcher die relative Verteilung amyloidophiler or amyloidophober Zustände des Monomers bestimmt, die Nachahmung vieler Phänomologien zu, was Aufschluss über die Kinetik der Aggregation geben kann.

Proteinfaltung ist ein Phänomen dem die Funktionalität von molekularen Komponenten zugrundeliegt. Vorhersagen welche Sequenzmodifikationen eine stabilere Struktur und möglicherweise ein funktionaleres Protein zur Folge haben, haben eine enorme Auswirkung auf molekül-basierte Technologien. Eine Zusammenarbeit mit Experimentalisten führte zur Entdeckung von Mutationen, die den gefalteten Zustand eines Proteins, bestehend aus mehreren Armadilloeinheiten, stabilisieren, das vorgängig Charakteristika eines geschmolzenen Kügelchens (*molten globule*) aufwies. Die rechenbetonten Beiträge halfen den Sequenzraum einzuschränken und die aussichtsreichsten Mutanten auszuwählen.

Membran-Protein Interaktionen bilden die Basis der Zellkommunikation und des Kerntransports. Diese Mechanismen, obwohl in einer Vielzahl von entscheidenden Prozessen involviert, sind nach wie vor auf molekularer Ebene schlecht verstanden. Die Erforschung von Peptiden mittels Simulationen der Moleküldynamik, die sowohl mit Membranen als auch mit Mizellen interagieren, verhalf zu neuen Einsichten dieser molekularen Mechanismen. Im ersten System konnte die spontane Faltung von Melittin an der Lipidoberfläche einer Mizelle reproduziert werden, währenddem die Gleichgewichtseigenschaften des Mizellen-Melittin-Komplexes ebenfalls erhalten blieb. In einer zweiten Arbeit wurden das in einer Membrandoppelschicht eingebettete, Lipid-modifizierte C-terminale Heptapeptid des menschlichen N-ras Proteins untersucht. Die Ergebnisse der Simulation bestätigten ein vorgängiges strukturelles Modell, basierend auf spektroskopischen Daten und schlugen einen Mechanismus des Einschubs der Peptids in die Membran vor.

Contents

3 Introduction

3.1 Open challenges in protein structural biology

Protein Folding. Proteins are ubiquitous and highly specialized macromolecules, which participate to nearly all metabolic process of the cell. They are polymerized by ribosomes in the endoplasmic reticulum as a linear chain of amino acids, whose sequence is encoded into the DNA. Due to intrinsic flexibility of the amino acids main chain, the ”backbone”, and the side-chains, a protein can assume a huge number of isomerization states. Nevertheless for many proteins only one specific isomerization is the functional ”native” conformation. Thus, once synthesized the protein must undergo to a complex process of folding to attain the functional isoform. In the cell, which is a highly crowded environment because of a concentration of 350 mg/ml [1] of macromolecules, the protein folding is sometimes assisted by various molecular helpers.

Energy landscape. Protein folding is not merely associated to the biological environment. Mechanistic investigation of protein folding have been advanced by the ability to observe these complex reactions also in vitro. Previously expressed and purified polypeptides can fold under physiological thermodynamic conditions also without assisting molecules: the protein folding to the native conformation is therefore a property of the single polypeptide chain. Folding or unfolding can be detected by the acquisition or loss of an enzymatic activity (denaturation), or a structural property inherent to the native state. Once completely denatured, a number of proteins can efficiently refold upon dilution from denaturant (refolding). Anfinsen in the late 50s has first monitored the activity of ribonuclease A upon denaturation [2], detecting the presence of a thermodynamics equilibrium between the folded and the unfolded state. Furthermore, he established that the necessary information to proper fold is univocally contained into the primary sequence. These earlier studies brought to the unanimously agreed picture that the native state of a protein is a minimum into the free energy landscape, where for landscape it is meant an abstract space consisting of all possible isomerization of the polypeptide chain.

Folding Thermodynamics and Kinetics. The thermodynamics of folding process is

determined by the free energy difference between the folded (F) and the unfolded (U) states ΔG. Though the protein folding is a reaction that can involve two or more states, in many cases it may be described by a simple two state reaction:

$$F \underset{k_f}{\overset{k_u}{\rightleftharpoons}} U \tag{1}$$

where k_u and k_f are the rate constants of the unfolding and folding reactions respectively. ΔG is derived from the equilibrium constant $K = k_f/k_u$ of the unfolding reaction:

$$\Delta G = -k_B T \log K \tag{2}$$

being k_B the Boltzmann constant and T the temperature in Kelvin. The presence of proteins that display an equilibrium in the folding-unfolding reaction with a pure two state kinetics, allowed the understanding of the process at a more theoretical and general level. An interesting feature of the protein folding is that it is not a sequential sampling of the isomerization space. In fact the estimated time to accomplish a random search is much longer than typical time scales of biological processes, thus useless for the scope of life [3]. Therefore, the folding process must be driven by a number of intermediates that efficiently restraints the isomerization space in a funnel-like manner. This suggests that not only the folding-unfolding thermodynamics, but also the kinetics is presculpted into the sequence.

Molten globule and intrinsically disordered proteins. Not all sequences have a stable folded structure under physiological conditions. For instance, molten globule is a particular state where a protein has a detectable secondary structure but misses of defined tertiary contacts. Furthermore as many as 30% of eukaryotic proteins are either completely disordered or owning disordered regions [4]. These proteins, called intrinsically disordered proteins (IDP) though not displaying any folded conformation, have indeed important roles into the cell, and represent an extention to the concept of functional conformation. IDPs have been found be involved in DNA/RNA-protein interaction, function as inhibitors or scavengers, and facilite the formation and function of multiprotein com-

plexes [5–7]. The discussion about molten globule will be continued in section 7, where an approach to the hydrophobic core stabilization will be presented.

Protein misfolding and aggregation. In contrast with the two state folding of model proteins, refolding from denaturant of large proteins is generally not efficient. For such proteins the fold is kinetically trapped in a local minima, and in this partly folded intermediate, protein can associate and eventually precipitate. A certain number of human diseases, classified as protein misfolding diseases, are due to an incorrect non-functional protein morphology competing with the native functional form (that could be either structured or unstructured). Recently attention has been focused on a group of misfolding diseases related to an abnormal deposit of polypeptide sequestered from their soluble form, which accumulate in different tissues [8]. Remarkably the main features of such deposits, known as amyloid deposits, are very similar among different precursor proteins, suggesting a common mechanism for the polymerization and, putatively, the toxic activity. In section 4 the phenomenology of amyloid aggregation is discussed more in detail, and in sections 5 and 6 the results of computational modelling of amyloid polymerization is presented.

Membrane protein. A large set of proteins belongs to the class of membrane proteins, which are polypeptides associated or integrally anchored to the cell membranes. These molecules play essential roles, such as signal transductions or mechanosensitive channels, and are the most important targets for drug discovery. Although it is estimated that 25-30 % of open reading frames encode for membrane proteins, only a few of their high-resolution structures have been resolved. And less is known about how membrane protein folds: it has been postulated that a two or three stages [9] mechanism is necessary for a complete α-helical transmembrane assembly, which might be assisted by molecular machinery such as translocons [10]. While structural characterization of large integral membrane molecules is still a big challenge, membrane associated polypeptides such as amphipathic helix or lipid modified peptides are within the range of standard structural investigation. In sections 8 and 9 molecular dynamics simulations are used to investigate the peptide-membrane interactions.

3.2 Computational chemistry

Computational chemistry is an established branch of scientific research whose contribution is determinant for the understanding of many biomolecular process. The molecular simulation, an applicative form of the statistical mechanics methods [11], and the concept of the *in silico* experiment are taking part to this wide discipline. In the *in silico* experiment a model of the real system is constructed, observables are measured and compared with real experimental properties. This comparison may validate or invalidate the model. If the model properties agree with real ones, the model can be used to make predictions. In particular computational chemistry is useful to describe:

- The structure and stability of a molecular system
- The free energy of different states of a molecular system
- Reaction process within molecular systems

During the decades, thanks to the improvement of informatics, the research enriched with new algorithms and new capabilities that yielded to an increase in complexity of the investigated systems. The field of computational chemistry experienced applications that ranged from the simple liquid phase transition to the protein folding, enzyme inhibition, membrane channel function, extending the impact of this field of research onto everyday life, with increasing level of reliability and prediction.

One of the most commonly used computational methods is Molecular Dynamics (MD), which is a technique that resolves the evolution of the degrees of freedom of molecular systems along the time. It can be roughly classified in two kinds: the *ab initio* methods, which applies the principles of quantum physics, and the molecular mechanics, which is based on the classical mechanics [12]. The latter uses an empirical energy potential that is called force field, and which is the set of prescriptions that define structure-energy relationship. Various force fields are currently being used, such as CHARMM [13], AMBER [14], GROMOS [15] and OPLS [16], and the typical equation for the potential energy is a pairwise, conservative function:

$$E(\mathbf{r}) = E_{cov}(\mathbf{r}) + E_{non-cov}(\mathbf{r}) \tag{3}$$

being E_{cov} and $E_{non-cov}$ the covalent and non-covalent energy terms, respectively. $\mathbf{r}$ is the $N \times 3$-dimensional array of atoms coordinates. The covalent term is the sum of bonds b, covalent angles θ, dihedral angles ϕ and improper dihedral angles ω:

$$\begin{aligned} E_{cov}(\mathbf{r}) &= E_b + E_\theta + E_\phi + E_\omega = \\ &= \sum_b \frac{1}{2} k_b (b - b_0)^2 + \sum_\theta \frac{1}{2} k_\theta (\theta - \theta_0)^2 + \\ &\quad \sum_\phi k_\phi [1 + \cos(n\phi + \psi)] + \sum_\omega k_\omega (\omega - \omega_0)^2 \end{aligned} \tag{4}$$

where k_b, k_θ, k_ϕ and k_ω are the force constants. Note that the bond, the angle and the improper potentials are harmonic potentials that assume the minimum value at r_{b_0}, θ_0 and ω_0. On the other side the torsional is a sum of periodic functions, whose indexes are n and phases are ψ.

The non-covalent term in equation (3) is the sum of Lennard-Jones and Coulombic potentials:

$$\begin{aligned} E_{non-cov}(r_{ij}) &= E_{LJ}(r_{ij}) + E_C(r_{ij}) \\ &= \sum_{i<j} \epsilon_{ij} \left[\left(\frac{r_{ij}^{min}}{r_{ij}} \right)^{12} - 2 \left(\frac{r_{ij}^{min}}{r_{ij}} \right)^{6} \right] + \sum_{i<j} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \end{aligned} \tag{5}$$

where q_i is the partial charge of atom i, ϵ_0 is vacuum permittivity, r_{ij}^{min} is the equilibrium separation distance of the Lennard-Jones potential and ϵ_{ij} is the energy well depth, i.e., $E_{LJ}(r_{ij}^{min}) = -\epsilon_{ij}$. The Lennard-Jones potential E_{LJ} is an effective pairwise function that mimics the van der Waals interaction, while the Coulombic potential E_C is the energy term that accounts for electrostatic attraction or repulsion between the charges. The trajectories of the atoms belonging to the molecular system are calculated by iterative numerical integration of the Newton equation of motion:

$$m_i \frac{d^2}{dt^2} \mathbf{r}_i = m_i \frac{d}{dt} \mathbf{v}_i = m_i \mathbf{a}_i = \mathbf{F}_i(\mathbf{r}_i) = -\nabla_i E(\mathbf{r}_i)$$

that expresses the relationship between the force field potential $E(\mathbf{r})$ defined in equation (3), and the kinematic quantities, i.e., the force F, the accelerations $\mathbf{a}$, the velocities $\mathbf{v}$ and the coordinates $\mathbf{r}$.

3.2.1 The solvation

Water, the main component of the biological liquid in which most macromolecules operate, is a complex medium, whose modellization is a particularly careful issue. A common approach is the use of explicit solvent, which consists in atomistic water molecules. This model has the advantage of being parameterizable at the water molecule level [17], though it has the major disadvantage to be extremely inefficient in the sampling of slow molecular events such as protein folding or aggregation. Another possible solution is the use of an implicit solvent, i.e., a mean field approximation of the solvent medium. In the implicit solvent models water degrees of freedom are not included into the force field potential but they are replaced by a potential that depends on the solute degrees of freedoms only [18]. The hydration of a molecule can be expressed by a reversible thermodynamic cycle where the free energy of solvation is the sum of polar and non-polar contributions:

$$G_{solv} = G_{np} + G_p = G_{cav} + G_{vdW} + G_p$$

where G_{cav} is the solvent cavitation term, G_{vdW} is the solute-solvent van der Waals and G_p is the solute-solvent electrostatic polarization term. Several implicit solvent schemes have been adopted to approximate either the non polar or the polar contributions. Continuum electrostatics approaches are the most reliable in reproducing the polar contributions, and they are based on the Poisson equation for the electrostatic potential $\psi(\mathbf{r})$ [19]:

$$\nabla \cdot [\epsilon(\mathbf{r}) \nabla \psi(\mathbf{r})] = -4\pi\rho(\mathbf{r})$$

where $\mathbf{r}$ is the vector of the solute atoms coordinates, ϵ is the position dependent dielectric and $\rho(\mathbf{r})$ is the fixed charge density of the solute. Numerical solution of Poisson equations have been implemented into the force fields [20], but are still very inefficient.

Hence, semi-analytical treatments of continuum electrostatics have been developed, including Generalized Born approximations [21], which reduce the electrostatic solvation energy to a pairwise potential. However the most efficient solvation schemes for molecular dynamics of macromolecules are those based on the solvent accessible surface area (SASA) [22–24] or slight variation of the same concept [25]. These models are based on the idea that solvation energies can be decomposed into atomic contributions, which are determined by the local solute geometry:

$$G_{solv} = \sum_i \sigma_i S_i \tag{6}$$

where i is the solute atom index, σ_i are surface tensions, and $S(i)$ are the atomic solvent accessible areas. The solute surface, which is a measure of the water accessibility in the first shell of hydration, is responsible for the main contributions to the solvation energies. This assumption holds for non-polar contributions, where it corresponds to the first order approximation of the scaled particle theory of solutes [26, 27], and has been observed in energetics of n-alkane transfer in water and amino acids [28]. The hydration Gibbs energy of proteins can be decomposed in chemical groups contributions [29], or atom type contributions [22]. In the latter case every atom type has its own surface tensions σ_i, which are derived from gas to water transfer energies of small compounds. Furthermore the charge screening operated by the solvent is approximated by a distant dependent dielectric with neutralized ionic groups [25].

3.2.2 The sampling

Molecular dynamics simulations generate a set of configurations of the system that rigorously obeys to the thermodynamic restraints imposed on the system (the statistical ensemble [11]). This process of coordinate exploration is called sampling and it is the basis for calculating the average value of an observable. Constant temperature simulations is a common way to perform a sampling. At fixed temperature, the probability that a physical system adopts a particular conformation $\mathbf{r}$ (with $\mathbf{r}_1, ..., \mathbf{r}_N$ coordinates of the

atoms) is given by the Boltzmann distribution at temperature T:

$$p(\mathbf{r}) = e^{-E(\mathbf{r})/kT} \tag{7}$$

where $E(\mathbf{r})$ is the potential energy of the conformation $\mathbf{r}$. The value associated to an observable A is then obtained as the weighted average over all possible conformations:

$$A = \sum_{\mathbf{r}} A(\mathbf{r}) p(\mathbf{r})$$

In a simulation, the measure of a physical quantity is simply as an average of the various instantaneous values assumed by the quantity during the MD run:

$$A = \frac{1}{S} \sum_{s} A(\mathbf{s})$$

where S is the total number of sampled conformations, and s is the index of conformation. Whether the generated set of conformations is representative for measuring observables depends on the extent to which the configuration space have been sampled, and depends on the sampling algorithm that is applied. In fact the time required to overcome a free energy barrier $\Delta G^{\ddagger}$ is:

$$\tau = \tau_0 \exp(\Delta G^{\ddagger}/k_B T)$$

with a prefactor τ_0 that has a typical value of 10^{-7} s. If we assume that for reactions such as folding, binding and aggregation the estimated barrier is about 10 kT, then the event has a single occurrence within the ms timescale [30]. These timescales are challenging the computational efficiency of modern computers, therefore strategies that improve the sampling have been developed.

Implicit solvent, as mentioned above, is a method that enhances the efficiency of calculation. Furthermore it is a way to improve also the sampling. In fact the viscosity effects due to the aqueous environment are neglected, reducing the overall frustration of the system.

Another way that allows improvement of both efficiency and sampling is the coarse graining [31]. Here the atomistic details are eliminated, and integrated into mesoscopic

units which interacts through a potential of mean force. The reduction of the number of interacting particles reflects into a dramatic gain of efficiency, but on the other side the atomistic details are lost. Coarse grained models are particularly attractive when the investigator is interested in studying phenomenologies at the mesoscopic scale, e.g., membrane-viruses interactions [32].

The reduction of the total number of the degrees of freedom of a system might help to improve efficiency. This method is particularly useful when the investigator is interested in a specific region of the molecule, e.g., the binding pocket of an enzyme or the hydrophobic core of a protein. By restraining or constraining the coordinates of the atoms not being part to the interested region one can substantially decrease the computation load. Furthermore some algorithms has been developed to satisfy bond geometry constraints during molecular dynamics simulations [33]. If applied to hydrogens, they allow a longer timestep integration.

Conformational sampling is a problem of considerable concern in the case of systems out of equilibrium,e.g., the amyloid fibril formation. In this case the state of the system evolves irreversibly towards a steady state, and the simulation of a single replica doesn't sample correctly the intermediates. A simple way to overcome this problem is to perform multiple molecular dynamics simulations [34] of the same system. Setting different initial velocities, the replicas evolve through completely different pathways and independently explore different intermediates. If the number of replicas is sufficiently large, the population of intermediates converges to the expected value. This method has been adopted by distributed computing [35] to exploit the world wide computational power of personal computers.

Among all algorithms, replica exchange methods, REM, is a very efficient way to simulate complex systems at low temperature. Sugita and Okamoto have extended the original formulation into an MD based version (REMD) [36], a technique that has been employed for studying the folding and aggregation of peptides [37–39]. The basic idea of REMD is to simulate different replicas of the system at the same time but at different temperature values. Each replica evolves independently by MD and periodically states

section of thesis	research project	force field/ solvent model	approximations	sampling
5	Amyloid aggregation rate prediction	empirical formula	based on protein sequence	not necessary
6	Amyloid aggregation mechanisms	coarse grained potential/ vacuum	monomers have 1 degree of freedom	multiple MD simulations
7	*de novo* protein core design	atomistic/ vacuum	restrained backbone	random rotamer sampling
8	Folding of helical peptide on micelle	atomistic solute and lipids/ implicit solvent	aqueous solvent is a continuum	constant temperature MD, REMD
9	Binding of lipidated peptide on membrane	atomistic solute and lipids/ explicit solvent	none	constant temperature MD, steered MD

Table 1: Biological problems and computational approaches.

with neighbor temperatures are swapped with a probability that is proportional to the potential energy difference between the two states. This produces a random walk in the space of temperature, which enables the overcome of barriers and improves low temperature sampling.

3.2.3 Multiscale modelling

Hence, the main problem of molecular simulation is how to sample efficiently the conformational space of molecules with a sufficiently accurate force field and solvation model. The relationship between accuracy and efficiency must be established by the requirements of the investigated system: assumptions, approximations and simplifications of the model must be chosen such that their contributions are comparable to the overall inaccuracy.

In this thesis application of computational chemistry to different biochemical problems is adapted to the degree of simplification. The chapters of this thesis are organized in increasing complexity and sophistication of the applied models. The sampling tech-

niques mentioned above have been employed as needed, as reported in Table 1.

In sections 5 and 6 two possible approaches are adopted to investigate and predict the kinetics of amyloid aggregation. The first method is an empirical analytical formula, and the second is a coarse-grained model of amyloid peptide simulated with molecular dynamics. In section 7 the optimization of the hydrophobic core of a designed protein is discussed. Section 8 the folding simulation of melittin into micelles with an implicit solvent model is treated. The final section concerns the explicit solvent simulation of a lipidated peptide binding onto a membrane bilayer.

4 Amyloid aggregation

The amyloid aggregates are fibrillar protein assemblies composed of polypeptide chains that are arranged in a β-conformation, with their backbone perpendicular to the axis of the fibril [40]. These aggregates, found in abnormal tissue deposits of several degenerative diseases, are the products of a complex oligomerization and polymerization cascade [41], and are related with the onset of cell dysfunction [8]. Amyloid accumulations are found in diseased brains: the amyloid plaques in the case of Alzheimer disease, the Lewy bodies in Parkinson disease, and the nuclear inclusions in Huntington disease. However, although in Alzheimer disease the presence of β-amyloid plaques in the central nervous system is associated with neurodegeneration and dementia [8], the fibrillar deposits themselves have not been proven to be directly related to the disease phenotypes [42]. Compelling evidences relate the toxicity of amyloid aggregation to the small oligomers which are released during the polymerization. Mixtures of soluble globular oligomers, which ranges from trimers to 24mers, have neurotoxic properties [43] and lead to fibril formation. Furthermore accumulation of soluble oligomers was verified in Alzheimer diseased frontal cortex [44] and were found to disrupt learning and interfere with cognitive functions in rats [45].

Atomic force microscopy and electron microscopy imaging revealed that *in vitro* amyloid fibrils are straight and unbranched, with a diameter which ranges between 1 and 10 nm, and they present a substructure consisting of multiple protofilaments twisted around the fibril axis. These aggregates also display a characteristic X-ray diffraction pattern with cross-β peaks, CD and FTIR spectra rich in β content, and a core structure resistant to hydrogen exchange and proteinase K digestion. The ability of dyes such as thioflavin T and Congo red to bind selectively to amyloid aggregates allows time resolved measurements of the fibrillization process, which helped to define the distinct kinetic phases. Amyloid fibril formation is a nucleated polymerization process (see figure 1), strongly modulated by both external conditions and polypeptide sequence. This process has a particularly complex phenomenology that presents a rich variety of intermediates [46, 47]. Furthermore the nucleation step can be bypassed inoculating a so-

lution of freshly dissolved monomers with seeds, which are fibrils that were previously sonicated. Some post-nuclear oligomers display a different morphology from mature fibrils. These ”protofibrils” are non-spherical filamentous aggregates lacking of periodicity that are observed to be either on or off-pathway to the formation of mature fibrils.

Experiments have shown that many proteins have amyloidogenic properties [48–50] including a number of intrinsically unstructured proteins [7]. Natively folded proteins can undergo to amyloid formation at low pH [51], high temperature [52, 53], moderate alcohol concentration [48], but also upon mutation [54]. Furthermore fibril formation is enhanced by denaturing conditions [55], suggesting that partial unfolding is a prerequisite for amyloid aggregation.

A number of experimental techniques have been employed to evaluate the β-aggregation propensity of polypeptides, i.e., the relative efficiency of a given sequence to form amyloid aggregates under given experimental conditions. Measurements based on kinetics [56] or thermodynamics [57] of the polymerization process can be used to quantify the β-aggregation propensity of a polypeptide. Interestingly, it is likely that small regions of the polypeptide sequence are responsible for the amyloidogenic behavior of the full length protein [58, 59]. Therefore these small regions, called often susceptible regions, or ”hot spots”, carry most of the amyloid propensity of the sequence.

Open issues regarding amyloid fibril formation include the specificity with which the amino acid sequence determines β-aggregation propensity, the atomic details of the fibril structure and the underlying molecular mechanisms. Investigations at the molecular level are helpful in this regard, but few structural models are currently available because of the difficulties encountered in X-ray crystallography and solution phase NMR spectroscopy. Computational approaches for predicting polypeptide aggregation propensities, as for instance the models discussed in section 5, are useful tools for determining aggregation hot spots and predicting effects of mutagenesis on the fibril formation kinetics [60]. However, details such as oligomer formation, lag phase time, pathways and nucleus size are not reproducible by such models. For that reason a coarse grained polypeptide under different conditions of concentration and aggregation propensity has been simulated to

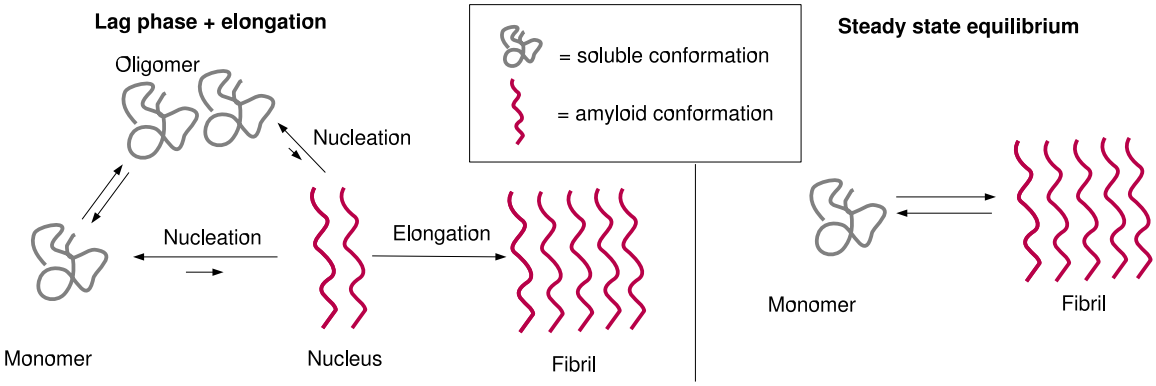


Figure 1: Amyloid formation displays the characteristic kinetics of a nucleated polymerization where the rate limiting step for fibril formation is the creation of a nucleus. In the figure the kinetic phases of fibril formation are illustrated. Left: nucleation (or lag phase) and elongation. In the lag phase monomeric or oligomeric soluble species are in pre-equilibrium with nuclei, which are unstable species that can either progress to fibril or regress to soluble species. Right: steady state equilibrium at the end of fibrillization.

investigate the effects of these agents on the nucleation velocity, the elongation and the pathways, as described in section 6.

5 Amyloid aggregation rate prediction

A systematic investigation of the effects over amyloid formation kinetics of 40 single point mutations of acyphosphatase elucidated the role of certain sequence regions in the unfolding aggregation mechanism [61], indicating that the most influential mutations (in terms of kinetics) are those that affects the propensity to form β-sheets and the hydrophobicity of the protein. Analysis of 17 Aβ40 mutants at the Val18 position illustrates the role of the mutation at the lag phase and the elongation [56]. Also in this case kinetics is influenced by β-sheet propensity, hydrophobicity. Chiti et al. [62] first hypothesized that kinetic data belonging to different protein mutants can be predicted by an empirical formula that involves difference in hydrophobicity $\Delta hydr$, free energy propensities $\Delta\Delta G_{coil-\alpha}$ and $\Delta\Delta G_{\beta-coil}$ and charge $\Delta Charge$:

$$ln(\nu_{mut}/\nu_{wt}) = A\Delta hydr + B(\Delta\Delta G_{coil-\alpha} + \Delta\Delta G_{\beta-coil}) + C\Delta Charge$$

where ν_{mut}/ν_{wt} is the ratio between the mutated and the non mutated elongation rate, and A, B and C are coefficients to be determined.

A possible role of aromatic residues and $\pi - \pi$ stacking has been postulated for amyloid aggregation and inhibition [63]. More recently it has been demonstrated that hydrophobic stretches in the sequence of Aβ 42 promote aggregation, but the presence of aromatic sidechains can sensibly accelerate the fibril formation [64]. Given these informations, and basing on physicochemical properties of amino acids, a function which predicts the change of amyloid aggregation rates upon mutation as been proposed (see section 5.1):

$$\nu_{mut}/\nu_{wt} = \phi_h\phi_\beta\phi_a\phi_c$$

where the factor ϕ_h includes nonpolar and polar interactions and it is proportional to the ratio of polar, nonpolar accessible surfaces and dipole change upon mutation, ϕ_β accounts for the β-sheet propensity change, ϕ_a is the change of aromatic residues and ϕ_c is the change of number of charged residues in the sequence. The novelty with respect

to Chiti's formula is the explicit introduction of a term depending on the quantity of aromatic residues. The function is able to predict relative rates with a correlation of 85%, better than Chiti's function which performs at 76%, and noticeably without any parameter, in contrast with Chiti's formula that contains three parameters.

The model has been further improved, with a parameterless function able to predict absolute rate of aggregation (see section 5.2). Over a set of 90 experimental data points, the predicted rates correlate at 95%. The advantage of having an absolute rate prediction consists in the detection of amyloid prone segments of the amino acids sequence. In Fig. 2 the amyloid spectrum of the Abeta42 is showed. The peaks in the amyloid spectrum can be interpreted as the aggregation susceptible regions ("hot spot") of the sequence. Remarkably the found aggregation spot (LVFFA) agrees with many experimental evidences [58], and the second highest region (AIIGL) and (IGLMV) are consistent with solid state NMR [65, 66]. The function can also predict the parallel or antiparallel arrangement of the segment into the fibril.

Complete proteomes of nine eukariotes of different complexity were analyzed by using the absolute rate formula (see section 5.3). Each protein was fragmented into stretches of 5 amino acids and analyzed in term of β-aggregation propensity. Statistical analysis of the stretch decomposition of proteomes. Form *P. Tetraurelia* to *H. Sapiens*, it has been shown that proteomes of higher and more long-lived eukariotes contain fewer sequences with high β-aggregation propensity. Also, compared with random proteomes,

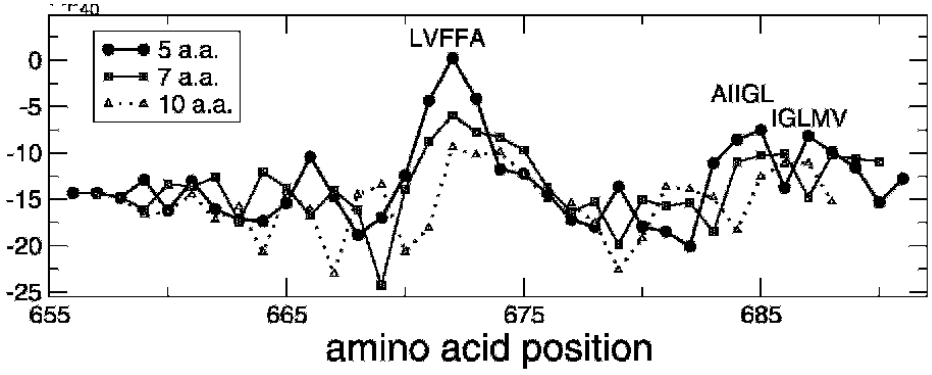


Figure 2: β-aggregation propensity profiles of the Alzheimer amyloid peptide Aβ42 (amyloid protein precursor numbering) evaluated with three different windows.

natural proteomes are enriched in proteins with low β-aggregation potential, as well as proteins with high β-aggregation potential. Such polarization is a consequence of the dual evolutive requirement of intrinsically disordered proteins with low β-aggregation propensity, as well as proteins with stable fold, which comes at the cost of higher β-aggregation propensity.

5.1 The role of aromaticity, exposed surface, and dipole moment in determining protein aggregation rates. [Protein Sci. 2004, *13*, 1939]

FOR THE RECORD

The role of aromaticity, exposed surface, and dipole moment in determining protein aggregation rates

GIAN GAETANO TARTAGLIA, ANDREA CAVALLI, RICCARDO PELLARIN, AND AMEDEO CAFLISCH

Department of Biochemistry, University of Zurich, CH-8057, Zurich, Switzerland

(RECEIVED February 2, 2004; FINAL REVISION March 18, 2004; ACCEPTED March 25, 2004)

Abstract

The mechanisms by which peptides and proteins form ordered aggregates are not well understood. Here we focus on the physicochemical properties of amino acids that favor ordered aggregation and suggest a parameter-free model that is able to predict the change of aggregation rates over a large set of natural sequences. Furthermore, the results of the parameter-free model correlate well with the aggregation propensities of a set of peptides designed by computer simulations.

Keywords: amyloid; prion; aggregation rate; Alzheimer; protein deposit; mutation

Supplemental material: see www.proteinscience.org

Amyloid fibrils are involved in a number of diseases, including Alzheimer's disease, Parkinson's disease, Huntington's disease, prion disease, and type II diabetes (Kelly 1998; Rochet and Lansbury Jr. 2000). Therefore, it is of fundamental medical interest to understand the mechanisms of fibrillogenesis with the ultimate goal of designing inhibitors. The amyloid fibril formation is not a property limited to a selected few proteins: Under certain conditions it has been shown that any polypeptide chain can form fibrils (Dobson 1999). Because aggregation conditions vary sensibly with the composition and sequence of the polypeptide, single amino acid substitution has been used to investigate the fibril formation (Chiti et al. 2002). In this study we propose a formula to predict the change of aggregation and disaggregation rate upon mutation. The agreement between the experimental data and our formula leads us to the conclusion that the formation of fibrils can be explained with a simple model based on physicochemical properties of amino acids. We found that the polar and the nonpolar water-accessible surface areas, the dipole moment, and the π-stacking interaction of aromatic residues (Gazit 2002) are essential beside the charge and the β-propensity of the sequence (Chiti et al. 2003). To have the most possible general model, we do not use any parameter that needs to be experimentally estimated. Furthermore, our equation does not present any redundancy, whereas in previous work by others charge and hydrophobicity were considered independent and used as two different variables in the best-fitting (Chiti et al. 2003).

We propose the following function to predict the effect of a mutation on aggregation rate:

$$\nu_{mut}/\nu_{wt} = \phi_h \phi_\beta \phi_a \phi_c \quad (1)$$

where ν_{wt} and ν_{mut} are the aggregation rates of the wild type and mutant, respectively. The factor ϕ_h captures most of the nonpolar and polar interactions. An amino acid is called *p* if its side chain carries a charge or a dipole; otherwise it is called *a*.

For mutations that involve same type of amino acids $a \rightarrow a$ or $p \rightarrow p$

$$\phi_h^I = \begin{cases} ASA^a_{mut}/ASA^a_{wt} & a \rightarrow a \\ ASA^p_{wt}/ASA^p_{mut} & p \rightarrow p \end{cases}$$

Reprint requests to: Amedeo Caflisch, Department of Biochemistry, University of Zurich, CH-8057, Zurich, Switzerland; e-mail: caflisch@bioc.unizh.ch; fax: +41-44-635-68-62.

Article published online ahead of print. Article and publication date are at http://www.proteinscience.org/cgi/doi/10.1110/ps.04663504.

where ASA^a and ASA^p are the nonpolar and polar water-accessible surface areas of the amino acid side chains (Makhatadze and Privalov 1990; Karplus 1997). Interestingly, experimental evidence has been published recently on the importance of nonpolar solvent-accessible surface area for the amyloid-like properties of apomyoglobin (Chow et al. 2003).

For mutations that involve different types of amino acids ($a \rightarrow p$ or $p \rightarrow a$)

$$\phi_h^{II} = \begin{cases} 1/D_{mut} & a \rightarrow p \\ D_{wt} & p \rightarrow a \end{cases}$$

where D is the magnitude of the dipole of the amino acid side chains. The function ϕ_h^I implies that the hydrophobicity and aggregation rate increase as the mutation results in a larger nonpolar surface or smaller polar surface. In ϕ_h^{II}, it has been assumed that the nonpolar surface of p amino acids compensates the nonpolar surfaces of a amino acids so that the dipole of p amino acids exclusively characterizes the mutation (see Supplementary Table 1).

The factor ϕ_β is related to the ratio of β-sheet propensities (Street and Mayo 1999; see Supplementary Table 1):

$$\phi_\beta = \frac{\beta_{mut}}{\beta_{wt}}$$

Functions ϕ_a and ϕ_c approximate the effect of the aromatic residues A and total charge C, respectively:

$$\phi_a \phi_c = e^{\Delta A} e^{-\Delta |C|/2}$$

The factor ½ before C has been introduced to have the same range [−1, 1] for the arguments of the two exponential functions.

In Figure 1 our model is used to predict the changes in aggregation rates occurring in human muscle acylphosphatase (AcP), islet amyloid polypeptide, prion peptides, α-synuclein, amyloid β-peptide, tau, leucine-rich repeat, and some model peptides. As in Chiti et al. (2003), we divided the data set in two parts to compare with their equation. The correlation obtained with equation 1 is significant (85% and 86% and $P < 10^{-4}$), and slightly better than the one obtained by Chiti et al. using three parameters derived from best fitting (76% and 85% and $P < 10^{-4}$). The good agreement with experiments shows that our simple equation, which does not contain any parameter, is very general and can be used to describe the aggregation of several and heterogeneous protein systems.

The validity of the formula is proved also by rearranging the whole data set per a and p mutations: Slopes and correlations are very close (see Supplementary Fig. 1; $p \rightarrow p$: slope = 1.01, correlation = 80%, number of points = 28; $a \rightarrow a$: slope = 0.92, correlation = 82%, number of points = 15; $a \rightarrow p$ and $p \rightarrow a$: slope = 1.01, correlation = 89%, number of points = 12).

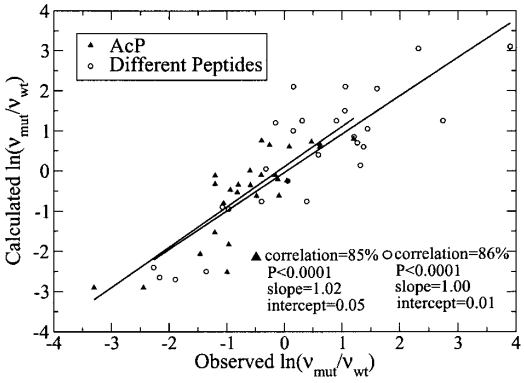


Figure 1. Calculated vs. observed (Chiti et al. 2003) changes in aggregation rate upon mutation: AcP (28 triangles) and heterogeneous groups of peptide and protein systems, including islet amyloid polypeptide, prion peptides, α-synuclein, amyloid β-peptide, τ, leucine-rich repeat and some model peptides (27 circles).

Aggregation and disaggregation are intrinsically different, but the role played by the hydrophobicity, β-propensity, π-stacking, and charge is the same. Considering that disaggregation and aggregation are opposite processes, the direct proportionality relation between ν_{mut}/ν_{wt} and $\phi_h\phi_\beta\phi_a\phi_c$ that describes the aggregation turns into a relation of inverse proportionality for the disaggregation. Therefore, the reciprocal of equation 1 can be used to describe the disaggregation:

$$\nu_{wt}/\nu_{mut} = \phi_h\phi_\beta\phi_a\phi_c \quad (2)$$

To verify the validity of this assumption, we applied equation 2 to heptapeptide sequences suggested by a genetic algorithm approach (G. Tartaglia and A. Caflisch, in prep.). The genetic algorithm searches the space of sequences for those that have the best match to a certain three-dimensional target conformation (an in-register parallel aggregate of three heptapeptides [Gsponer et al. 2003]). For each peptide sequence, three replicas are submitted to a 330 K molecular dynamics simulation, starting from the β-parallel aggregated conformation (CHARMM parameter 19 [Brooks et al. 1983] and solvent accessible surface-based solvation model [Ferrara et al. 2002]). A temperature of 330 K is used to obtain enough sampling in the time scale of the simulations (Gsponer et al. 2003). Peptide sequences are ranked according to their ability to prevent disaggregation. The disaggregation rate is estimated for each sequence as the reciprocal of the number of snapshots whose C_α root mean square deviation (RMSD) from the template is lower than 1 Å. Best

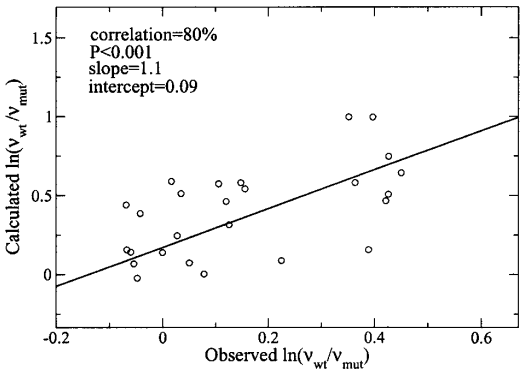


Figure 2. Calculated vs. observed changes in disaggregation rate upon mutation: Best parents of genetic algorithm approach (27 circles). (See Supplementary Table 2.)

matches, called best parents, are replicated and subjected to mutations and crossover: 10^3 sequences have been studied for a total amount of 50 μsec of simulation. The genetic algorithm predicted several sequences similar to segments of amyloidogenic protein as well as the sequence HFWLVFF, which presents five matches with the amyloid β-peptide fragment HQKLVFF (Tjernberg et al. 1999; Williams et al. 2004). By considering that the genetic algorithm sampled 10^3 sequences and a random search approximately needs 10^6 sequences to scan before finding five matches, we conclude that the genetic algorithm approach performs 10^3 better than random.

Disaggregation rates are analyzed with equation 2 only for best parents (4% of data) for which false positives are supposed to be less than the false negatives in the remaining set. Furthermore, to have statistical significance, each disaggregation rate has been averaged over a set of five molecular dynamics trajectories. Figure 2 shows that equation 2 holds and the correlation is very high (80% and $P < 10^{-3}$). In conclusion, the present results indicate that a simple model based on physicochemical properties without parametrization is able to predict aggregation and disaggregation rates.

Acknowledgments

We thank Dr. E. Paci for interesting discussions and Prof. F. Chiti for providing rates of AcP. This work was supported by the Swiss National Science Foundation (grant no. 31-64968.01 to A.C.) and the National Center of Competence in Research (NCCR) in Structural Biology.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

References

Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S., and Karplus, M. 1983. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **4:** 187–217.

Chiti, F., Calamai, M., Taddei, N., Stefani, M., Ramponi, G., and Dobson, C. 2002. Studies of the aggregation of mutant proteins in vitro provide insights into the genetics of amyloid diseases. *Proc. Natl. Acad. Sci.* **99:** 16419–16426.

Chiti, F., Stefani, M., Taddei, N., Ramponi, G., and Dobson, C. 2003. Rationalization of the effects of mutations on peptide and protein aggregation rates. *Nature* **424:** 805–808.

Chow, C., Chow, C., Raghunathan, V., Huppert, T.J., Kimball, E.B., and Cavagnero, S. 2003. Chain length dependence of apomyoglobin folding: Structural evolution from misfolded sheets to native helices. *Biochemistry* **42:** 7090–7099.

Dobson, C.M. 1999. Protein misfolding, evolution and disease. *Trends Biochem. Sci.* **24:** 329–332.

Ferrara, P., Apostolakis, J., and Caflisch, A. 2002. Evaluation of a fast implicit solvent model for molecular dynamics simulations. *Proteins* **46:** 24–33.

Gazit, E. 2002. A possible role for π-stacking in the self-assembly of amyloid fibrils. *FASEB J.* **16:** 77–83.

Gsponer, J., Haberthür, U., and Caflisch, A. 2003. The role of side-chain interactions in the early steps of aggregation: Molecular dynamics simulations of an amyloid-forming peptide from the yeast prion Sup35. *Proc. Natl. Acad. Sci.* **100:** 5154–5159.

Karplus, P. 1997. Hydrophobicity regained. *Protein Sci.* **6:** 1302–1307.

Kelly, J. 1998. The alternative conformations of amyloidogenic proteins and their multi-step assembly pathways. *Curr. Opin. Struct. Biol.* **8:** 101–106.

Makhatadze, G. and Privalov, P. 1990. Heat capacity of proteins 1 partial molar heat capacity of individual amino acid residues in aqueous solution: Hydration effect. *J. Mol. Biol.* **213:** 375–384.

Rochet, J.C. and Lansbury Jr., P.T. 2000. Amyloid fibrillogenesis: Themes and variations. *Curr. Opin. Struct. Biol.* **10:** 60–68.

Street, A. and Mayo, S. 1999. Intrinsic β-sheet propensities result from van der Waals interactions between side chains and the local backbone. *Proc. Natl. Acad. Sci.* **96:** 9074–9076.

Tjernberg, L., Callaway, D., Tjernberg, A., Hahne, S., Lilliehook, C., Terenius, L., Thyberg, J., and Nordstedt, C. 1999. A molecular model of Alzheimer amyloid β-peptide fibril formation. *J. Biol. Chem.* **274:** 12619–12625.

Williams, A., Portelius, E., Kheterpal, I., Guo, J., Cook, K., Xu, Y., and Wetzel, R. 2004. Mapping A_β amyloid fibril secondary structure using scanning proline mutagenesis. *J. Mol. Biol.* **335:** 833–842.

5.2 Prediction of aggregation rate and aggregation-prone segments in polypeptide sequences. [Protein Sci. 2005, *14*, 2723]

Prediction of aggregation rate and aggregation-prone segments in polypeptide sequences

GIAN GAETANO TARTAGLIA, ANDREA CAVALLI, RICCARDO PELLARIN, AND AMEDEO CAFLISCH

Department of Biochemistry, University of Zürich, CH-8057 Zürich, Switzerland

(RECEIVED March 23, 2005; FINAL REVISION June 23, 2005; ACCEPTED July 4, 2005)

Abstract

The reliable identification of β-aggregating stretches in protein sequences is essential for the development of therapeutic agents for Alzheimer's and Parkinson's diseases, as well as other pathological conditions associated with protein deposition. Here, a model based on physicochemical properties and computational design of β-aggregating peptide sequences is shown to be able to predict the aggregation rate over a large set of natural polypeptide sequences. Furthermore, the model identifies aggregation-prone fragments within proteins and predicts the parallel or anti-parallel β-sheet organization in fibrils. The model recognizes different β-aggregating segments in mammalian and nonmammalian prion proteins, providing insights into the species barrier for the transmission of the prion disease.

Keywords: Alzheimer's disease; amyloid; protein aggregation rate; prion protein; species barrier; genetic algorithm; molecular dynamics

Amyloid fibrils are associated with a number of pathologies including Alzheimer's, Parkinson's, Huntington's, prion disease, and type II diabetes (Horwich and Weissman 1997; Kelly 1998; Dobson 1999; Rochet and Lansbury 2000). Therefore it is of fundamental medical interest to understand the mechanisms of fibrillogenesis, with the ultimate goal of designing inhibitors. One important and still unanswered question regarding amyloid fibril formation is the specificity with which the amino acid sequence determines β-aggregation propensity and the atomic details of the fibril structure. Because of the difficulties in obtaining detailed structural information by X-ray crystallography or solution phase NMR spectroscopy, computational approaches are needed to guide experiments, e.g., to determine short segments of amyloid-like proteins that share the same biophysical properties of the full-length proteins (Balbirnie et al. 2001) and identify those elements which are essential for the formation of protein fibrils (Tenidis et al. 2000; von Bergen et al. 2000). As aggregation conditions vary sensibly with the composition and especially the sequence of the polypeptide, single amino acid substitutions have been used to investigate the fibril formation (Chiti et al. 1999), and complementary theoretical studies proposed relative rate equations to predict the change of aggregation rate upon mutation (Chiti et al. 2003; Tartaglia et al. 2004). Although the application of relative rate equations shows high correlation with experimental data, these models require the a priori knowledge of wild-type aggregation rates.

We report here an absolute rate equation derived from both first principles and analysis of aggregating sequences designed by a computational approach. The latter is based on a genetic algorithm optimization in sequence space and molecular dynamics sampling of conformation space. The equation does not need any information except the amino acid sequence and two environmental factors (i.e., temperature and concentration). Our model gives both the aggregation rate and the "amyloid spectrum" of a protein, identifying those segments involved in β-aggregation. In

Reprint requests to: Amedeo Caflisch, Department of Biochemistry, University of Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland; e-mail: caflisch@bioc.unizh.ch; fax: +41-44-635-68-62.

Article and publication are at http://www.proteinscience.org/cgi/doi/10.1110/ps.051471205.

addition, the model distinguishes between the parallel and anti-parallel β-sheet organization within the fibrils and shows that mammalian and nonmammalian prion proteins have different amyloid spectra.

Results and Discussion

Absolute rate prediction

Predicted and experimentally measured rates are shown in logarithmic scale in Figure 1. The correlation is 95% and extends over 90 data points and about 15 natural logarithmic units. This is a remarkable result considering that the rate is calculated solely from the primary structure with the addition of two external factors, i.e., temperature and concentration. Interestingly, the correlation is good for different proteins and also within mutants of the same protein. For single-point mutants of long sequences (Acylphosphatase and Titin), the error is rather large because of the poor signal-to-noise ratio due to the average over the entire sequence. The model was subjected to statistical tests to assess the chance correlation. In Figure 2A, the experimentally measured rates were randomly permutated to generate about 10^7 "scrambled" data sets. The calculated rates were fitted to each scrambled set, giving an extremely small likelihood for high correlations. In Figure 2B, $\sim 10^7$ data sets were randomly generated within the range of experimental rates. The predictive ability and correlation of the model are much higher than the corresponding values obtained upon randomization of the experimental rates. These statistical tests show that chance correlation is not present.

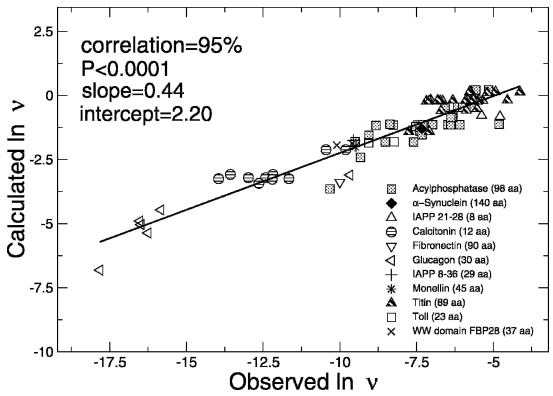


Figure 1. Calculated (Equation 4; see Materials and Methods) vs. observed aggregation rates for heterogeneous groups of peptide and protein systems (Litvinovich et al. 1998; Konno et al. 1999; Chiti et al. 2003; Ferguson et al. 2003; DuBay et al. 2004). A t-student test on the correlation shows the high significance in the prediction (in the present study $P < 0.0001$, while $P \simeq 1$ indicates no significance).

Prediction of β-aggregating segments

There is in vivo evidence that amyloid fibrils originate from misfunctions of the degradation machinery and cleavage of fragments that have high propensity for β-aggregation (Stefani and Dobson 2003). Moreover, even proteins not implicated in amyloid diseases were recently found to form amyloid fibrils in vitro under denaturing conditions, indicating that fibrillogenesis is a common feature of proteins (Chiti et al. 1999; Dobson 1999; Stefani and Dobson 2003). Our approach to estimate aggregation rates can be also used to identify segments with high aggregation propensity. The method is tested on the following proteins: α-synuclein, apolipoprotein, amyloid precursor protein (APP), gelsolin, islet amyloid precursor protein (IAPP), lactadherin, prion, serum amyloid A, transthyretin, ABri, ADan, fibrinogen, β_2-microglobulin, insulin, Sup35, and tau. The former nine proteins represent all hits of a combined search for "amyloid" and "human" at http://www.expasy.org (Gasteiger et al. 2003) in September 2004; the latter seven proteins result from a literature search (references are reported in Table 1). As indicated in Figure 3, the data set contains

- regions known to promote aggregation
- segments found to aggregate in vivo (often after degradation)
- stretches extracted from the precursors and shown to aggregate in vitro

Each sequence in the data set is scanned by shifting a window of fixed size one residue at a time starting from the N terminus. The extracted stretches are ranked using the aggregation propensity π (see Materials and Methods). The procedure is repeated for different window sizes (3–25 amino acids), each time storing the positions of the three stretches having the highest π. These positions are then used to build the histogram of Figure 3. Peaks of the histogram represent positions of stretches with the highest β-aggregation propensity ("windows' consensus"). All the sequences except fibrinogen and prion show main peaks in segments known to promote aggregation. For prion, amyloidogenic areas are—up to now—not known and few experiments have been performed and on limited portions of the protein (Vanik et al. 2004). Following the protein-only hypothesis (Prusiner 1988; Soto and Castilla 2004), we suggest that the peak found at position 150 may be determinant for prion transmissions (in the subsection Prions, the same peak is numbered with 175 because of the alignment with other prion sequences). For transthyretin, only one of the two experimentally known β-aggregating fragments has been found with our analysis. We speculate that the corresponding area promotes the aggregation of the entire protein, which is consistent with NMR data (Jaroniec et al. 2002).

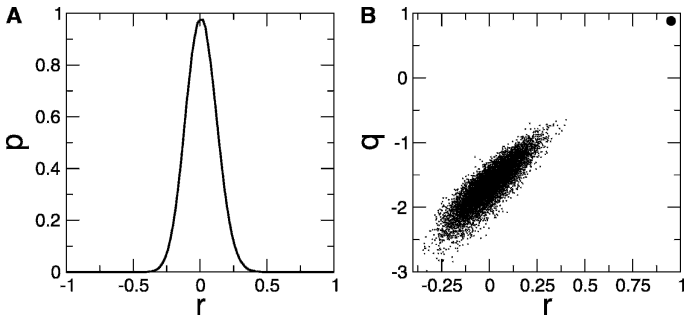


Figure 2. Statistical tests to assess chance correlation. (*A*) Permutations of experimental rates: Probability distribution p of the correlation coefficient r between rates calculated with Equation 4 (see Materials and Methods) and scrambled experimental rates. The likelihood of obtaining high correlations ($r > 50\%$) with scrambled experimental rates is extremely small ($p < 10^{-9}$). (*B*) Randomization of experimental rates (within the same range of values): Cross-validated leave-one-out correlation coefficient $q = 1 - \text{PRESS}/\sigma^2$ (PRESS = predicted residual sum of squares, i.e., sum of squared differences between predicted and observed values [Zoete et al. 2003]) vs. the correlation coefficient r. The predictive ability and correlation of the model (thick circle on the *top right*) are significantly separated from the corresponding values obtained upon randomization of the experimental rates (thin points). In both tests, 10^7 data sets were generated.

To further test the sensitivity of our model, we focused on the segments that are experimentally known to aggregate. For this purpose, we used a window size of five consecutive residues, as in a previous work (Fernandez Escamilla et al. 2004) (Table 1). Interestingly, several five-residue stretches are found in segments that were shown to aggregate, e.g., FGAIL contained in IAPP NFGAILSS, FILDL in gelsolin's SFNNGDCFILD, SVQFV in lacthaderin's NFGSVQFV, and YQQYN in Sup35's PQGGYQQYN (Azriel and Gazit 2001). For APP, three stretches are found in correspondence of the segment LVFFA, which is known to be involved in the aggregation of $A\beta_{40}$ (Williams et al. 2004) (see subsection Amyloid Protein Precursor). Importantly, all the stretches are ranked among those having the

Table 1. *Analysis of experimentally known* β*-aggregating segments*

Protein	1st Stretch[a]	Rank[b]	2nd Stretch[a]	Rank[b]	3rd Stretch[a]	Rank[b]	Segment	Total length	Ref.
ABri	22{CSRTV}^{a}	5	21{ICRST}^{a}	8	20{LICSR}^{a}	10	1–34	34	El-Agnaf et al. 2001
ADan	22{CFLNF}^{p}	1	23{FNLFL}^{p}	2	24{NLFLN}^{p}	3	1–34	34	El-Agnaf et al. 2004
α-Synuclein	41{EQVTN}^{a}	6	67[SIAAA]^p	12	71[ATGFV]^p	15	41–74	120	Ueda et al. 1993
Apolipoprotein A–I	18[YVDVL]^p	1	28{DYVSQ}^{a}	2	85[EMSKD]^a	3	1–83	242	Nichols et al. 1988
APP	671{LVFFA}^{p}	1	670{KLVFF}^{p}	2	672{VFFAE}^{p}	3	655–696	750	Weidemann et al. 1989
β-Microglobulin	61{SFYLL}^{p}	1	63{TLLYY}^{p}	2	66{YYTEF}^{p}	3	59–79	99	Jones et al. 2003
Fibrinogen	494[FPGFF]^p	7	493[TFPGF]^p	13	482[AAFFD]^p	32	482–504	623	Asl et al. 1997
Gelsolin	187{DCFIL}^{p}	15	188{CFILD}^{p}	23	189{FILDL}^{p}	31	173–243	755	Kangas et al. 1996
IAPP	22{FGAIL}^{p}	1	21{NFGAI}^{p}	2	28[SNTYG]^a	4	1–38	38	Westermark et al. 1987
Insulin	78{ENYCN}^{a}	1	23{RGFFY}^{p}	3	15{ALYLV}^{p}	4	1–38	86	Jimenez et al. 2002
Lactadherin	260{YGNDQ}^{a}	3	259{SYGND}^{a}	4	289[SVQFV]^p	5	245–294	364	Haggqvist et al. 1999
Prion	116{IIHFG}^{p}	1	115{PIIHF}^{p}	2	99[VVGGL]^p	3	1–121	208	Vanik et al. 2004
Serum amyloid A	3{FFSFL}^{p}	2	4{FSFLG}^{p}	3	5{SFLGE}^{p}	4	2–12	104	Westermark et al. 1992
Sup35	77[YQQYN]^a	1	44[YQNYQ]^a	2	67[YQQQY]^a	3	1–112	683	King et al. 1997
Tau	621{SVQIV}^{p}	23	632{SKVTS}^{a}	24	627{KPVDL}^{p}	25	617–636	757	Margittai and Langen 2004
Transthyretin	107[IAALL]^p	1	114{YSYST}^{a}	2	106[TIAAL]^p	4	105–115	127	Jaroniec et al. 2002

[a] The three five-residue stretches with the highest π, within the segments listed in the third to last column, are reported with the predicted parallel (*p*) or anti-parallel (*a*) arrangement. The braces { } indicate stretches that are close to the peak found in the experimental regions using the windows' consensus (Figure 3), while the brackets [ ] mark sequences that are distant from the peak. The integer before the brackets refers to the position of the stretch in the processed protein (initial signal- and pro-peptides are omitted in the notation as in other works; see, for instance Kangas et al. 1996; Jones et al. 2003).

[b] The rank of the stretches refers to the entire precursor protein and can in principle vary from 1 (i.e., the stretch has the highest π among all the stretches in the precursor protein) to the total length of the precursor protein (i.e., the stretch has the lowest π among all the stretches in the precursor protein).

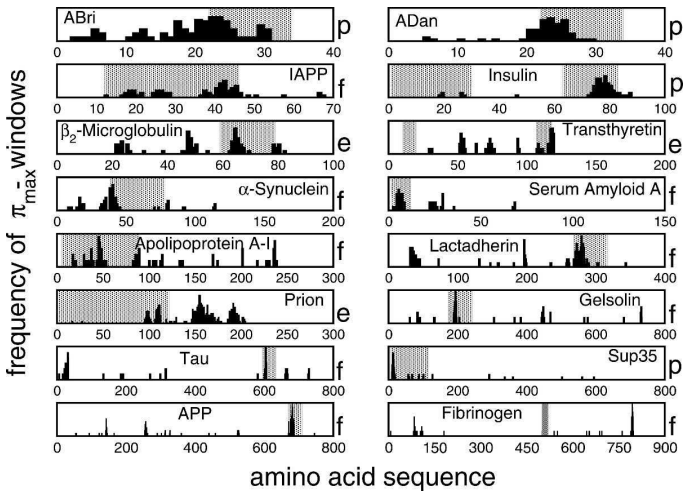


Figure 3. Windows' consensus. Different window sizes (3–25 amino acids) are used to scan proteins. Positions of stretches with highest aggregation propensity π are used to build the histogram. Except for fibrinogen and prion, the highest peak is located in segments that are known to form amyloid fibrils and/or contribute to protein aggregation (gray regions). The letter "p" labels regions that are known to promote fibrillogenesis ("p" standing for "promoting"). The letter "f" indicates segments that are found to aggregate in vivo ("f" standing for "fragment") after degradation. The letter "e" refers to stretches that are shown to aggregate in vitro ("e" standing for "extracted"). We stress that Equation 1 (see Materials and Methods) was used to identify β-aggregating stretches and not to predict amino acid deletions or insertions involved in amyloidosis. Positions refer to proteins without signal- and pro-peptides. References for all the experiments are reported in Table 1.

highest π in the respective precursor proteins (see Table 1), which suggests that a small window size is sufficient for the identification of amyloidogenic regions. In Table 1, we also list β-aggregating segments that have not yet been investigated with experiments in vitro (e.g., YVDVL in apolipoprotein A–I and ENYCN in insulin) and indicate the predicted parallel or anti-parallel arrangement of the individual segments in the fibril.

Amyloid protein precursor

Using a window size of five residues, the amyloid spectrum of the 750-residue APP (Fig. 4) shows a predominant peak at position 671 for the stretch LVFFA. Furthermore, the predicted β-aggregating stretches AIIGL and IGLMV are consistent with solid-state NMR (Antzutkin et al. 2002; Bond et al. 2003) and scanning proline mutagenesis (Williams et al. 2004). The stretches with the highest rate for each window size in the range 3–25 are shown in Table 2 for $A\beta_{42}$. Most of the high-aggregation stretches contain the segment LVFFA and are parallel. As in experiments (Gordon et al. 2004), the segment KLVFFAE has a preferential anti-parallel arrangement, while $A\beta_{42}$ is parallel (Antzutkin et al. 2000; Torok et al. 2002). As shown in clinical reports and oligomerization experiments performed with photo-induced cross-linking of unmodified proteins (Bitan et al. 2003), we found that $A\beta_{42}$ has a higher aggregation propensity than $A\beta_{40}$ ($\ln \pi_{A\beta_{42}} = -7$, $\ln \pi_{A\beta_{40}} = -9$). Interestingly, the experimental evidence indicates that the Ile_{41}–Ala_{42} extension of the 1–40 segment affects the rate of amyloid formation rather than the fibril stability (Jarrett et al. 1993).

Prions

To further investigate the usefulness of our model, the amyloidogenic propensities of the prion protein from different organisms were evaluated using a moving window of five residues along the entire sequence. To compare the amyloid spectra, prion sequences have been aligned using ClustalW (Thompson et al. 1994). It is remarkable that prion sequences in mammals show a peak at position 175 corresponding to the segment SNQNN in human prion (Fig. 5; Table 3; all the notations used to number stretches refer to the major prion proteins, i.e., signal- and/or propeptides are omitted). Such a peak is absent in the chicken and the turtle. Interestingly, the peak is located in a glutamine/asparagine-rich region, which shows high propensity to self-propagate in amyloid fibrils (Michelitsch and Weissman 2000). Other peaks correspond to β-strand 2 (segment NQVYY, conserved in mammals and nonmammals and mutated in NRVYY in chicken) and helix 1 of

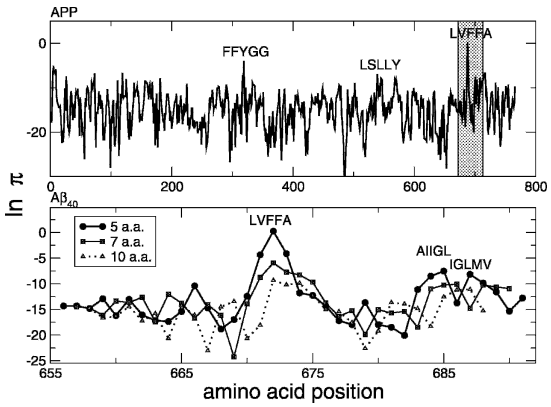


Figure 4. Amyloid protein precursor. The aggregation propensity π is averaged over a window of five amino acids. The entire sequence is scanned by shifting the window by one residue at a time starting from the N terminus ("amyloid spectrum"). The analysis shows a major peak corresponding to the segment LVFFA at position 671. The *bottom* plot focuses on the most amyloidogenic region, which is highlighted in gray in the *top* plot. Windows of different sizes (5, 7, and 10 amino acids), shifted to the central amino acid, give similar results, indicating the robustness of the model. Furthermore, with longer window sizes, peaks in the C terminus of $A\beta_{40}$ become comparable to the one at position 671 (see also Table 2). In both plots, the effective height of the peak is compressed by the logarithm scale.

human prion (segment YEDRY in mammals, WNENS in turtle, and WSENS in chicken), which are known to form ordered aggregates in vitro (Nguyen et al. 1995; Kozin et al. 2001). Furthermore, the amyloid profiles are similar within mammals (e.g., 97% correlation between man and cow) and different between mammals and nonmammals (e.g., 55% correlation between man and turtle).

To compare with experiments in vitro (Vanik et al. 2004), we analyzed the unstructured region of the prion protein (residues 1–122) in human, mouse, and hamster prion peptides. We found that human and mouse prions share similar amyloid spectra (i.e., 98% correlation), while the hamster prion diverges from them at position 143 (position 116 in the nonaligned human sequence). More specifically, the stretch 143–148 of hamster prion (position 116–121 in the nonaligned human sequence) is found to be less amyloidogenic than the corresponding segment in mouse and human ($\ln \pi_{hamster} = -16$, $\ln \pi_{mouse} = -12$, and $\ln \pi_{human} = -12$), which is consistent with the prion 1–122 species barrier observed in vitro (Vanik et al. 2004).

Huntingtin

The gene for Huntington's disease consists of 67 hexons and contains an open reading frame for a polypeptide of > 3140 residues. Using a window size of five residues, our model identifies the N-terminal poly(Gln) repeat and the stretch IFFFL in the middle of the sequence as the two most prone to induce ordered aggregates. With window sizes larger than 20, the N-terminal poly(Gln) repeat dominates and the peak in the middle of the sequence disappears.

Our model is not sensitive enough to discriminate repeats of fewer than 38 glutamine residues from those with > 41 glutamine residues; the former are harmless, whereas the latter are responsible for toxic aggregates (Perutz et al. 1994; Perutz 1999). Alternatively, the dramatic difference in toxicity observed at a repeat length of ~40 might require the context of a much longer polypeptide sequence.

Conclusions

The model presented here was motivated by the challenging tasks of predicting aggregation propensity and identifying β-aggregating stretches in polypeptide sequences. An essential element in the derivation of the equation was the analysis of a large pool of β-aggregating peptide sequences designed by a computational approach based on molecular dynamics and genetic algorithm optimization in sequence space (G.G. Tartaglia and A. Caflisch, in prep.). The very

Table 2. *Stretches of $A\beta_{42}$ with the highest rate at each window size in the range 3–25*

Sequence		ln π	p/a
VFF	{IGL}	5.3 {−2.6}	p
LVFF	{GAII}	2.5 {−6.7}	p
LVFFA	{AIIGL}	0.2 {−7.5}	p
LVFFAE	{GAIIGL}	−3.9 {−8.0}	p
KLVFFAE	{AIIGLMV}	−5.9 {−10.0}	a
LVFFAEDV	{IGLMVGGM}	−7.3 {−10.1}	p
LVFFAEDVG	{GLMVGGVVI}	−7.6 {−10.0}	p
QKLVFFAEDV	{IGLMVGGVVI}	−9.3 {−9.7}	a
{QKLVFFAEDVG}	IGLMVGGVVIA	−10.1 {−11.0}	p
{HQKLVFFAEDVG}	AIIGLMVGGVVI	−10.5 {−11.1}	p
{FFAEDV . . . }	GAIIGLMVGGVVI	−10.5 {−10.7}	p
FFAEDVGSNKGAII		−10.1	p
VFFAEDVGSNKGAII		−9.3	p
VFFAEDVGSNKGAIIG		−9.7	p
LVFFAEDVGSNKGAIIG		−8.8	p
LVFFAEDVGSNKGAIIGL		−8.2	p
KLVFFAEDVGSNKGAIIGL		−9.3	p
KLVFFAEDVGSNKGAIIGLM		−9.4	p
QKLVFFAEDVGSNKGAIIGLM		−10.5	p
QKLVFFAEDVGSNKGAIIGLMV		−10.1	p
LVFFAEDVGSNKGAIIGLMVGGV		−10.7	p
LVFFAEDVGSNKGAIIGLMVGGVV		−10.4	p
LVFFAEDVGSNKGAIIGLMVGGVVI		−7.1	p

In braces are reported stretches that ranked after the highest rate ones and do not overlap with them. The last column contains the preferred β-sheet arrangement, i.e., parallel (p) or anti-parallel (a).

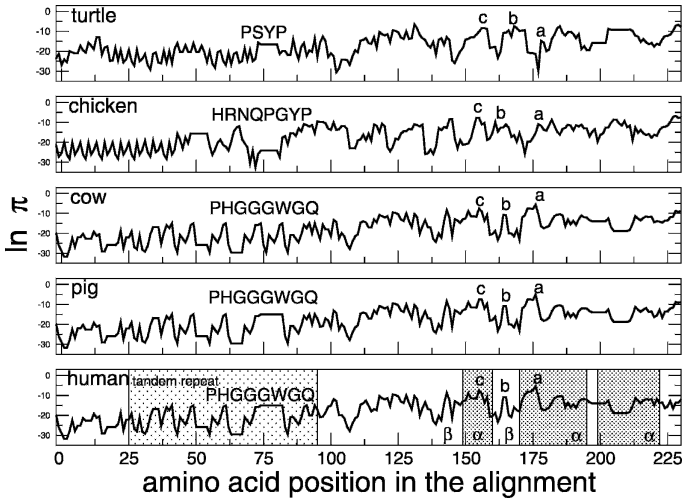


Figure 5. Prion proteins from turtle to human. The plot shows an evolutionary differentiation of the aggregation peaks. Prions of cow and mouse, as well as prions of sheep and pig, show similar amyloid spectra (data not shown). The highest peak at position 175 for mammals (segment *a*, i.e., SNQNN) is not present in nonmammals. Peak *b* (segment NQVYY, conserved in mammals and nonmammals, and mutated to NRVYY in chicken) appears in correspondence of β-strand 2 in human prion. Nonmammals show a peak *c* (segment WNENS in turtle and WSENS in chicken) in correspondence of the first helix of human prion that is weaker in mammals (YEDRY). Sequences have been aligned using ClustalW (Thompson et al. 1994) at http://www.expasy.org/cgi-bin/hub (Gasteiger et al. 2003). Horizontal traits in the plots represent gaps and are meant to help the eye. For all the species, no significant peak is found in the N-terminal tandem repeats. The secondary structural elements of the human prion are labeled with Greek letters and the stretches corresponding to the three α-helices are emphasized by shadowed rectangles.

good correlation between calculated and experimental rates for a large and heterogeneous set of polypeptide chains has allowed us to use the model to successfully identify β-aggregating segments and predict the parallel or anti-parallel arrangement. Fibrils formed by short segments of a protein might have a different molecular structure than the fibril of the full-length protein. Yet our results, as well as previous experimental (Chiti et al. 1999, 2003; Balbirnie et al. 2001) and computational (Fernandez Escamilla et al. 2004) works by others, indicate that the amyloid-forming part of a protein could be only a short segment of the entire chain. That a function based on simple physicochemical principles is able to predict aggregation rates and identify β-aggregating fragments in proteins might be a consequence of the essential role of side-chain interactions in β-sheet aggregates (Gazit 2002; Gsponer et al. 2003; Linding et al. 2004).

Although some of the physicochemical properties in our model are similar to those used in previous works by others, it is important to distinguish approaches based on parameter optimization for a multiterm equation (Chiti et al. 2003; DuBay et al. 2004) from first-principle models like the one of this work and that of Tartaglia et al. (2004). On a very similar test set of peptides and proteins, the multi-parameter approach gives results comparable to those obtained with our model, but it is likely to have a lower predictive ability. As an example, positional effects are taken into account in our model, whereas they are neglected in the multiparameter approach (DuBay et al. 2004), which is mainly based on amino acid composition and alternation of hydrophobic–hydrophilic residues (Broome and Hecht 2000). Recent scanning proline mutagenesis, combined with critical concentration analysis and NMR hydrogen–deu-

Table 3. *Peak at position 175. Prion compatibilies of animals with respect to human*

Animal	$\Delta\pi/\pi$
Turtle	9.52
Chicken	8.72
Sheep	1.66
Pig	1.13
Cow	0.76
Mouse	0.76
Hamster	0.15

The distance with respect to the human prion sequence is measured as $\Delta\pi/\pi = (\pi_{animal} - \pi_{human})/\pi_{human}$ using a window size of five amino acids for the rate calculation and summing over the segment 165–185 to better sample the variability around the peak.

terium exchange, indicate a strong positional effect on both the aggregation kinetics and structural properties of the $A\beta_{40}$ fibril (Williams et al. 2004). Most importantly, the multiparameter approach cannot be used to identify β-aggregating segments as explicitly mentioned by the investigators (DuBay et al. 2004).

Recently, an approach based on secondary structure propensity and estimation of desolvation penalty (TANGO) has been shown to accurately predict the sequence-dependent and mutational effects on the aggregation of a large data set of peptides and proteins (Fernandez Escamilla et al. 2004). TANGO is based on the assumption that the probability of finding > 2 ordered segments in the same polypeptide is negligible. The investigators report that TANGO allows quantitative comparison within the same polypeptide chain or mutants. On the other hand, only qualitative comparison between different polypeptide chains is possible with TANGO (Fernandez Escamilla et al. 2004), whereas our model allows for the prediction of absolute rates (Fig. 1).

In conclusion, we have identified the physicochemical properties of amino acids that are essential for ordered aggregation and proposed a model that takes into account sequence effects for aromatic and charged residues, as well as composition. Compared with the models previously published by others, our equation is the only one that takes explicitly into account π-stacking. Very recent high-resolution structural data (electron and X-ray diffraction) have provided strong evidence for the importance of aromatic side chains for amyloid formation (Makin et al. 2005).

Our model derived from first principles and analysis of *in silico* designed sequences is able to predict aggregation rates and identify β-aggregating segments with high accuracy, suggesting possible biological implications as in the prion protein case. For nonmammalian prions, the absence of the peak at position 175 observed in mammals decreases the overall aggregation propensity, indicating a species-specific behavior consistent with experiments (Marcotte and Eisenberg 1999; Matthews and Cooke 2003) and supporting the hypothesis of a species barrier in the transmission of the prion disease (Hill et al. 2000).

In the accompanying article we present a bioinformatics application of our model that reveals an anti-correlation between organism complexity and proteomic β-aggregation propensity (Tartaglia et al. 2005).

Materials and methods

Absolute rate equation

An equation based on physicochemical properties of natural amino acids is introduced to estimate the aggregation rate of proteins and identify β-aggregating segments. Aromaticity, β-propensity, and formal charges play a major role in our model, as they are known in the literature to be determinant for fibrillization (Gazit 2002; Tjernberg et al. 2002; Chiti et al. 2003). Polar and nonpolar surfaces, as well as solubility, are also taken into account following an analysis of sequences designed to aggregate into β-sheets. The design of β-aggregating sequences was performed by structural sampling using molecular dynamics and peptide sequence optimization by a genetic algorithm (Tartaglia et al. 2004; G.G. Tartaglia and A. Caflisch, in prep.) (see subsection Derivation of the Equation). The aggregation propensity π_{il} of an l-residue segment starting at position i in the sequence is evaluated as:

$$\pi_{il} = \phi_{il}\Phi_{il} \tag{1}$$

The factor Φ_{il} contains exponential functions and is position-dependent

$$\Phi_{il} = e^{A_{il}+B_{il}+C_{il}} \tag{2}$$

where A_{il}, B_{il}, and C_{il} are functionals related to the aromaticity, β-propensity, and charge, respectively. The factor ϕ_{il} depends almost exclusively on the amino acid composition

$$\phi_{il} = \left[\prod_{j=i}^{i+l-1}\left(\frac{S_j^a}{\hat{S}^a}\theta^{\uparrow\uparrow} + \frac{S_j^p}{\hat{S}^p}\theta^{\uparrow\downarrow}\right)\frac{\hat{S}^t}{S_j^t}\frac{\hat{\sigma}}{\sigma_j}\right]^{1/l} \tag{3}$$

where S_i^a, S_i^p, S_i^t, and σ_i—weighted by their average over the 20 standard amino acids (hatted values)—are the side-chain apolar, polar, total water-accessible surface area, and solubility, respectively (see subsection Parallel and Anti-Parallel Configuration). The functionals $\theta^{\uparrow\uparrow}$ and $\theta^{\uparrow\downarrow}$ include positional effects and reflect the parallel or anti-parallel tendency to aggregate if the majority of residues is apolar or polar, respectively. Considering the high correlation between measured and predicted changes in aggregation rate upon single point mutations (Chiti et al. 2003; DuBay et al. 2004; Tartaglia et al. 2004), it is possible to utilize the propensity π_{il} to predict the absolute rate v_{il}

$$\nu_{il} = \alpha(c,T)\ \pi_{il} \tag{4}$$

where $\alpha(c,T)$ is introduced to take into account concentration and temperature (see subsection Concentration and Temperature).

Parallel and anti-parallel configuration

The functional for the parallel or anti-parallel configuration was introduced following the analysis of sequences designed by genetic algorithm optimization (see subsection Derivation of the Equation; Fig. 6):

- The parallel in-register β-sheet organization within fibrils is favored by the number of side chains involved in π-stacking (Tyr, Phe, and Trp) and apolar interactions (Ala, Gly, Ile, Leu, Met, Pro, and Val) (McGaughey et al. 1998; Azriel and Gazit 2001; Jenkins and Pickersgill 2001; Makin et al. 2005). The number of aromatic and apolar residues is indicated with $n_{aromatic}$ and n_{apolar}, respectively. Hydrogen bonds

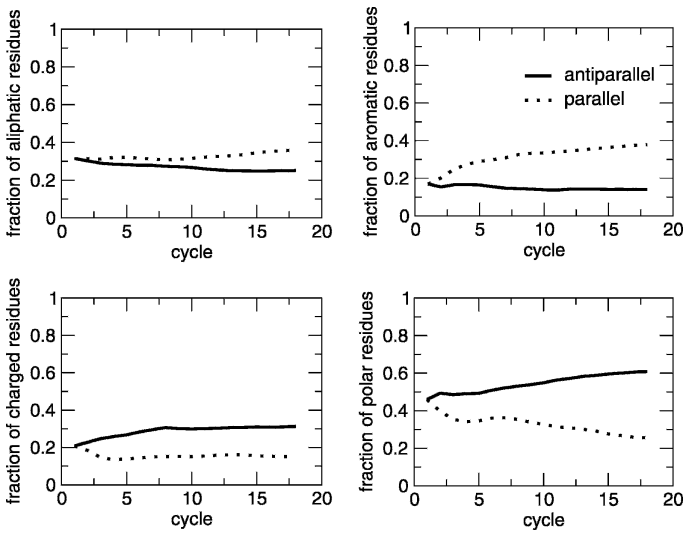


Figure 6. Computational design: A genetic algorithm approach was developed to search the space of peptide sequences for those with the best match to a given three-dimensional target conformation, i.e., an in-register parallel or anti-parallel aggregate of three heptapeptides (Gsponer et al. 2003). For each peptide sequence, three replicas were submitted to a 330 K molecular dynamics simulation, starting from the β-aggregated conformation using CHARMM parameter 19 and a solvent-accessible surface-based solvation model (Brooks et al. 1983; Ferrara et al. 2002). The sequence optimization was performed by evolutionary cycles. A total of 1728 sequences was sampled after 18 cycles. In sequences selected for the parallel aggregation, the number of aliphatic and aromatic residues increases almost monotonically, while the number of charged and polar residues decreases. The opposite is observed in sequences selected for the anti-parallel aggregation. In the plots, the number of aliphatic, aromatic, charged, and polar residues is normalized by the length of the peptide and averaged over the population (48 peptides per cycle).

between polar residues are not considered for the parallel aggregation because the number of polar residues decreases significantly during the optimization of parallel aggregated sequences (Fig. 6).

- The anti-parallel configuration is mainly determined by the electric dipole moment of the polypeptide (Hwang et al. 2004). Sequences abounding in polar residues show a small tendency for the parallel in-register aggregation because of unfavorable dipole–dipole interactions between side chains. Hence, the anti-parallel organization is promoted by the number of polar residues (Arg, Asn, Asp, Cys, Gln, Glu, His, Lys, Ser, and Thr), which is indicated with n_{polar}. In some specific positions, charged (Arg, Lys, Asp, and Glu) and aromatic amino acids contribute to the anti-parallel aggregation. "Specific positions" means that one or more couples of opposite charged residues or one or more aromatic residues are symmetrically placed with respect to the center of the sequence (Balbach et al. 2000; Hwang et al. 2004; Makin et al. 2005). In this specific case, the number of charged and aromatic residues is labeled as n^s_{charge} and $n^s_{aromatic}$, respectively.

In Equation 3, a parallel configuration is preferred if $n_{apolar} + n_{aromatic} > n_{polar} + n^s_{charge} + n^s_{aromatic}$. Since the number of aromatic residues in symmetric position is always smaller than the total amount of aromatic residues, $n_{aromatic} \geq n^s_{aromatic}$ (e.g., in the APP stretch: LVFFA $n_{aromatic} = 2$, $n^s_{aromatic} = 1$), we used a stricter condition for the parallel arrangement $n_{apolar} > n_{polar} + n^s_{charge}$. The stricter condition allows the factorization of aromatic contributions in Equation 1. In the ϕ_{il} factor of Equation 3, $\theta^{\uparrow\uparrow}$ and $\theta^{\uparrow\downarrow}$ are

$$\theta^{\uparrow\uparrow} = \begin{cases} 1 & n_{apolar} \geq n_{polar} + n^s_{charge} \\ 0 & otherwise \end{cases}$$
$$\theta^{\uparrow\downarrow} = 1 - \theta^{\uparrow\uparrow}$$

It is useful to explain the effect of the $\theta^{\uparrow\uparrow}$ and $\theta^{\uparrow\downarrow}$ functional by some examples. The segment LVFFA at position 671–676 of the APP is predicted to be parallel because it satisfies the parallel condition $n_{apolar} > n_{polar} + n^s_{charge}$ with $n_{apolar} = 3$ and $n_{polar} = n^s_{charge} = 0$ ($\theta^{\uparrow\uparrow} = 1$). The segment KLVFFAE (at position 670–677 of the APP), with two opposite charged residues, has anti-parallel propensity because it satisfies the anti-parallel condition $n_{apolar} < n_{polar} + n^s_{charge}$ with $n_{apolar} = 3$ and $n_{polar} = n^s_{charge} = 2$ ($\theta^{\uparrow\downarrow} = 1$).

The IAPP stretch FGAIL at position 22–26 is predicted to be parallel ($n_{apolar} = 4$ and $n_{polar} = n^s_{charge} = 0$, i.e.,$\theta^{\uparrow\uparrow} = 1$), in agreement with experimental results (Kayed et al. 1999a; Azriel and Gazit 2001; Gazit 2002). As in Azriel and Gazit (2001), the following stretches are predicted to be parallel: SVQFV at position 289–292 of lactadherin; DCFIL, CFILD,

and FILDL at position 187–191, 188–192, and 189–193 of gelsolin, respectively; FFSFL, FSFLG, and SFLGE at position 3–7, 4–8, and 5–9 of serum amyloid, respectively.

Poly(Gln), poly(Asn), and poly(Lys) homopolymers are predicted to be in an anti-parallel arrangement, as proposed in Perutz et al. (1994), Scherzinger et al. (1997), and Michelitsch and Weissman (2000) and observed by Tanaka et al. (2001) and Dzwolak et al. (2004). Moreover, it is likely that completely aliphatic sequences result in amorphous aggregates if N and C termini are capped, while a tendency to the anti-parallel arrangement is expected for short stretches with charged termini (e.g., transthyretin's stretch IAALL). Capping groups are neglected in the present version of the model.

The fragment GNNQQNY from the Sup35 yeast prion is predicted to be anti-parallel ($n_{apolar} = 1$, $n_{polar} = 5$, and $n^s_{charge} = 0$, i.e.,$\theta^{\uparrow\downarrow} = 1$), in contrast with the parallel packing suggested on the basis of X-ray diffraction and Fourier transform infrared (FTIR) data (Balbirnie et al. 2001). On one hand, it is important to note that the experimental data supporting a parallel arrangement are not conclusive, and, in particular, FTIR can be misleading on this point. In fact, in the unit cell of the microcrystals, the parallel β-sheets are proposed to be in anti-parallel contact along the fibril axis. On the other hand, a possible reason for the parallel configuration is that the π-interactions between the Tyr side chains are much less favorable in the anti-parallel configuration.

Aromatic residues

Aromatic side chains contribute to the parallel aggregation with π-interactions (McGaughey et al. 1998; Azriel and Gazit 2001; Makin et al. 2005). The density of aromatic residues $n_{aromatic}/l$ is used to distinguish two regimes for the aromatic contribution A_{il} of Equation 2:

$$A_{il} = \begin{cases} A^{low}_{il} & n_{aromatic}/l \le 3/20 \\ A^{high}_{il} & otherwise \end{cases}$$

where 3/20 is the aromatic density averaged over the 20 standard amino acids and $n_{aromatic}$ was defined in the previous subsection. In the case of low aromatic density ($n_{aromatic}/l \le 3/20$), A^{low}_{il} takes into account the polar/apolar environment. Following the results obtained by the genetic algorithm optimization of β-aggregation-prone sequences (see Fig. 6), A^{low}_{il} has a positive effect for mainly apolar sequences and a negative contribution for mainly polar sequences:

$$A^{low}_{il} = n_{aromatic}\,[n_{apolar} - (n_{polar} + n^s_{charge})]\; l^{-1}$$

The variables n_{apolar}, n_{polar}, and n^s_{charge} are defined in the previous subsection.

As an example, the APP stretch LVFFAEDVGSNKGAIIGLMVGGVVI shows low aromatic density ($n_{aromatic}/l = 2/25 < 3/20$). Since $i = 671$, $l = 25$, $n_{apolar} = 17$, $n_{polar} = 6$, and $n^s_{charge} = 0$, the aromatic contribution for LVFFAEDVGSNKGAIIGLMVGGVVI is $A^{low}_{671\ 25} = 2\ [17 - 6]\ 25^{-1} = 0.88$.

In the case of a high aromatic density ($n_{aromatic}/l > 3/20$), the model takes into account the number of aromatic residues:

$$A^{high}_{il} = n_{aromatic}$$

As an example, the APP stretch LVFFA shows high aromatic density ($n_{aromatic}/l = 2/5 > 3/20$). Since $i = 671$ and $l = 5$, the aromatic contribution for LVFFA is $A^{high}_{671\ 5} = 2$.

Besides the total amount of aromatic residues and the position dependence, which enters Equation 2 through A^{low}_{il}, the different polar and apolar side-chain surfaces, solubility, and β-propensity of Phe, Tyr, and Trp are taken into account in the factor ϕ_{il}. Hence, the mutation F22Y for the IAPP (islet β-amyloid protein precursor) stretch NFGAILSS produces a sensible change of rate ($\ln \pi_{wt} = -6, \ln \pi_{F22Y} = -7$), compatible with experiments in vitro (Porat et al. 2003).

β-Propensity

The β-propensity is evaluated as the fraction of residues that stabilize the β-sheet more than the α-helix:

$$B_{il} = \beta_{il}\, l^{-1} - 1/2$$

The function β_{il} is defined as:

$$\beta_{il} = \sum_{j=i}^{i+l-1} \delta^{\beta}_j$$

where

$$\delta^{\beta}_j = \begin{cases} 1 & \beta_j \ge \alpha_j \\ 0 & otherwise \end{cases}$$

The variables α_j and β_j correspond to the α-helix and β-sheet stabilizing effects of the amino acid at position j (Fersht 1999). Values of α_j and β_j are normalized from 0 (low stabilization) to 1 (high stabilization) to have the same range of variability. In the function B_{il}, the offset value of 1/2 is introduced so that $B_{il} > 0$ if at least one-half of the residues in the sequence is more stable in a β-sheet rather than in an α-helix conformation (i.e., $\beta_{il} > l^{-1}/2$).

In the case of the APP stretch LVFFA, values are $i = 671$, $l = 5$, $\beta_{672} = \beta_{673} = \beta_{674} = 1$, and $\beta_{671} = \beta_{675} = 0$. The predicted β-propensity for LVFFA is $\beta_{671\ 5} = 3/5 - 1/2 = 0.1$.

Charged residues

As in other models, we consider that the electrostatic repulsion of charged sequences penalizes the aggregation (Chiti et al. 2003; Tartaglia et al. 2004). In addition, our model takes into account the fact that side-chain pairs with opposite charges and positioned symmetrically with respect to the center of the segment contribute to the anti-parallel aggregation, as found in experiments (Gordon et al. 2004). In Equation 2, the charge contribution C_{il} is

$$C_{il} = -\frac{n_{charge}}{l}\left|\sum_{j=i}^{i+l-1} C_j\right| + \sum_{j=i}^{i+l-1} \delta^{charge}_j$$

where C_j is the charge of the side chain and n_{charge} is the number of charged residues. The first term of the functional C_{il} takes into account the electrostatic repulsion between polypeptides with net charge different from zero. The second term

counts the number of pairs of opposite charged side chains that are symmetrically placed with respect to the central residue of the sequence:

$$\delta_j^{charge} = \begin{cases} 1 & C_j = -C_{2i+l-j-1} \text{ and } C_j \neq 0 \\ 0 & otherwise \end{cases}$$

In the case of the APP stretch KLVFFAE, the residues K_{670} and E_{676} have opposite charges and are symmetrically placed with respect to the central amino acid F_{673}. Since $i = 670$, $l = 7$, $C_{670} = +1$, and $C_{1340+7-670-1} = C_{676} = -1$, the net charge for KLVFFAE is $|\sum_{j=i}^{i+l-1} C_j| = 0$ and the oppositely charged K_{670} and E_{676} give $C_{670\ 7} = \delta_{670}^{charge} + \delta_{676}^{charge} = 2$.

Surfaces and solubility

For sequences that are predominantly apolar ($\theta^{\uparrow\uparrow} = 1$; see subsection Parallel and Anti-Parallel Configuration), the apolar water-accessible surface S_j^a measures the contribution of hydrophobic side chains to aggregation. For mostly polar sequences ($\theta^{\uparrow\downarrow} = 1$), the polar water-accessible surface S_j^p takes into account the propensity to form hydrogen bonds between polar residues. The total surface $S_j^t = S_j^a + S_j^p$ is used to weight polar and apolar surfaces by the total area. Values of apolar and polar side-chain surfaces are given in our previous work (Tartaglia et al. 2004) and span the intervals 44–195 Å^2 and 27–107 Å^2, respectively. Averaged values are $\hat{S}^a = 108$ Å^2 and $\hat{S}^p = 54$ Å^2. In the case of poly(Gln), values of surfaces are $S^a = 53$ Å^2 and $S^p = 91$ Å^2. Since Gln is polar and $\theta^{\uparrow\downarrow} = 1$, the surface contribution is $S^p/\hat{S}^p \cdot \hat{S}^t/S^t = (91/54)(162/144) = 1.9$.

The variable σ_j takes into account the water solubility of the side chain at position j. In our model, aggregation propensity and solubility are inversely proportional to introduce a penalty for highly soluble polypeptides. Most of the solubility values are available at http://acrux.igh.cnrs.fr/proteomics/densities_pi.html (Nahway 1989). The missing values (Cys, Lys, and Thr) were taken from http://www.formedium.com/Europe/amino_acids_and_vitamins.htm. The variable σ_j spans the interval 0.04–162 g/100 g, with average $\hat{\sigma} = 3.95$ g/100 g. In the case of poly(Gln), $\hat{\sigma}/\sigma = 3.95/2.5 = 1.5$, which indicates low solubility in agreement with experiments of β-aggregation (Perutz et al. 1994; Perutz 1999).

Concentration and temperature

The function $\alpha(c,T)$ captures the effects of concentration (c) and temperature (T) in Equation 4:

$$\alpha(c,T) = RT \begin{cases} c & c \in [0, c^*]\ mM \\ 1 & c \in (c^*, 1]\ mM \\ 1/c & c > 1\ mM \end{cases}$$

The aggregation rate v is approximated to be proportional to the temperature because the probability of collision and elongation of peptides increases with temperature (Kusumoto et al. 1998). Although aggregation rate and temperature are not expected to correlate above physiological values (Massi and Straub 2001), we used a simple linear dependence, which is preferable for the small extent of experimentally accessible values of the temperature. In fact, the temperature ranges from 298 K to 310 K in the data set of Figure 1.

In agreement with quasielastic light-scattering experiments of fibrillogenesis of $A\beta_{40}$, the aggregation rate v is assumed to be proportional to the concentration for $c < c^*$ mM ($c^* = 0.1$ mM) and to be independent of concentration above the critical value $c = c^*$ (Lomakin et al. 1996, 1997) (see also subsection-Derivation of the Equation). The hyperbolic function $1/c$ was introduced to decrease the aggregation rate v for $c > 1$ mM, as there is experimental evidence that a very high concentration opposes formation of ordered aggregates (Munishkina et al. 2004). The concentration ranges from 0.01 mM to 20 mM in the data set of Figure 1.

Derivation of the equation

- Functionals for aromaticity, β-propensity, and charge were taken from our relative rate equation (Tartaglia et al. 2004). The aromatic term was modified according to the results obtained by the genetic algorithm optimization of aggregating sequences (Fig. 6) (G.G. Tartaglia and A. Caflisch, in prep.). The functional for β-propensity, previously based on a single scale (Tartaglia et al. 2004), now takes into account β- versus α-propensity. Scales for β- and α-propensity are taken from Fersht (1999) and normalized in the range 0–1. The term used for the β-propensity was tested on 100 globular proteins: 82% of the β-sheet content is successfully recognized (data not shown). The functional for charged residues was modified with the addition of a term for symmetrically placed charges of opposite signs, which is consistent with experimental data (Gordon et al. 2004). The function n_{charge}/l replaces the constant factor in the relative rate (Tartaglia et al. 2004) and is introduced to weight the overall charge by the charge density. The three functionals for aromaticity, charge, and β-propensity can be zero. Exponential functions were introduced so that their product is different from zero.
- The product of the three functionals was plotted versus available experimental rates (see next subsection), obtaining a correlation of 80%, while the individual correlations for aromaticity, charge, and β-propensity are 76%, 81%, and 70%, respectively.
- The dependence on concentration and temperature was introduced to derive aggregation rates from propensities (Lomakin et al. 1997; Kusumoto et al. 1998; Massi and Straub 2001; Munishkina et al. 2004). With the concentration alone, the correlation improves to 85%. The correlation is 82% without the hyperbolic function for high concentrations ($c > 1$ mM). With the temperature function, the correlation improves to 88%.
- The factor for polar/apolar contributions ϕ_{il} in Equation 1 was added upon the analysis of sequences produced by computational design (Fig. 6). The term is a linear combination of normalized surfaces and has a nonzero minimum. The correlation improves to 92%. The solubility dependence was added at the very end and introduces a penalty for highly soluble sequences. The correlation improves to 95%.

Experimental data

Most of the experimental rates were kindly provided by Dr. F. Chiti and Dr. M. Vendruscolo (Chiti et al. 2003; DuBay et al. 2004). The remainder data set was taken from previous experimental studies (Litvinovich et al. 1998; Konno et al. 1999; Ferguson et al. 2003). The absolute aggregation rates were

determined from in vitro experiments of denaturated polypeptide chains without taking into account the presence of cellular components as chaperones and proteases. Aggregation rates were obtained from kinetic traces in different ways: thioflavin T fluorescence, turbidity, CD, sedimentation, size exclusion chromatography, and filtration. Lag phases were not considered in the analysis, as they were not reported or difficult to extract from published data (DuBay et al. 2004). Since a comprehensive understanding of lag phases in protein aggregation is lacking (Kayed et al. 1999b; Padrick and Miranker 2002) (e.g., it is not known whether fibrils form by addition of monomers or oligomers and how growth conditions influence the amyloid formation), the aggregation kinetics was analyzed after the lag phase. The elongation phase showing an exponential behavior is fitted to the function $z = \alpha (1 - e^{-vt})$ where v is the rate measured in sec^{-1}.

Acknowledgments

We thank Prof. C. Dobson, Prof. F. Chiti, Dr. M. Vendruscolo, and Dr. J. Zurdo for providing rates of several proteins. The molecular dynamics simulations were performed on the Matterhorn Beowulf cluster at the Informatikdienste at the University of Zurich. We thank C. Bolliger, Dr. T. Steenbock, and Dr. A. Godknecht for setting up and maintaining the cluster. This work was supported by the Swiss National Science Foundation and the NCCR "Neural Plasticity and Repair."

The program for the calculation of aggregation rates is available from the corresponding author upon request.

References

Antzutkin, O.N., Balbach, J.J., Leapman, R.D., Rizzo, N.W., Reed, J., and Tycko, R. 2000. Multiple quantum solid-state NMR indicates a parallel, not antiparallel, organization of β-sheets in Alzheimer's β-amyloid fibrils. *Proc. Natl. Acad. Sci.* **97:** 13045–13050.

Antzutkin, O.N., Leapman, R.D., Balbach, J.J., and Tycko, R. 2002. Supramolecular structural constraints on Alzheimer's-amyloid fibrils from electron microscopy and solid-state nuclear magnetic resonance. *Biochemistry* **41:** 15436–15450.

Asl, L.H., Liepnieks, J.J., Uemichi, T., Rebibou, J.M., Justrabo, E., Droz, D., Mousson, J.M.C., Benson, M.D., Delpech, M., and Grateau, G. 1997. Renal amyloidosis with a frame shift mutation in fibrinogen α-chain gene producing a novel amyloid protein. *Blood* **90:** 4799–4805.

Azriel, R. and Gazit, E. 2001. Analysis of the minimal amyloid-forming fragment of the Islet amyloid polypeptide. *J. Biol. Chem.* **276:** 34156–34161.

Balbach, J.J., Ishii, Y., Antzutkin, O.N., Leapman, R.D., Rizzo, N.W., Dyda, F., Reed, J., and Tycko, R. 2000. Amyloid fibril formation by a β16–22, a seven-residue fragment of the Alzheimer's β-amyloid peptide, and structural characterization by solid state NMR. *Biochemistry* **39:** 13748–13759.

Balbirnie, M., Grothe, R., and Eisenberg, D. 2001. An amyloid-forming peptide from the yeast prion Sup35 reveals a dehydrated β-sheet structure for amyloid. *Proc. Natl. Acad. Sci.* **98:** 2375–2380.

Bitan, G., Kirkitadze, M.D., Lomakin, A., Vollers, S.S., Benedek, G.B., and Teplow, B.D. 2003. Amyloid Aβ-protein Aβ assembly: Aβ40 and Aβ42 oligomerize through distinct pathways. *Proc. Natl. Acad. Sci.* **100:** 330–335.

Bond, J.P., Deverin, S.P., Inouye, H., El-Agnaf, O.M.A, Teeter, M.M., and Kirschnera, D.A. 2003. Assemblies of Alzheimer's peptides Aβ25–35 and Aβ31–35: Reverse-turn conformation and side-chain interactions revealed by x-ray diffraction. *J. Struct. Biol.* **141:** 156–170.

Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S., and Karplus, M. 1983. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **4:** 187–217.

Broome, B.M. and Hecht, M.H. 2000. Nature disfavors sequences of alternating polar and non-polar amino acids: Implications for amyloidogenesis. *J. Mol. Biol.* **296:** 961–968.

Chiti, F., Calamai, M., Taddei, N., Stefani, M., Ramponi, G., and Dobson, C.M. 1999. Studies of the aggregation of mutant proteins in vitro provide insights into the genetics of amyloid diseases. *Proc. Natl. Acad. Sci.* **99:** 16419–16426.

Chiti, F., Stefani, M., Taddei, N., Ramponi, G., and Dobson, C.M. 2003. Rationalization of the effects of mutations on peptide and protein aggregation rates. *Nature* **424:** 805–808.

Dobson, C.M. 1999. Protein misfolding, evolution and disease. *Trends Biochem. Sci.* **24:** 329–332.

DuBay, K.F., Pawar, A.P., Chiti, F., Zurdo, J., Dobson, C.M., and Vendruscolo, M. 2004. Predicting absolute aggregation rates of amyloidogenic polypeptide chains. *J. Mol. Biol.* **341:** 1317–1326.

Dzwolak, W., Muraki, T., Kato, M., and Taniguchi, Y. 2004. Chain-length dependence of α-helix to β-sheet transition in polylysine: Model of protein aggregation studied by temperature-tuned FTIR spectroscopy. *Biopolymers* **73:** 463–469.

El-Agnaf, O.M.A., Sheridan, J.M., Sidera, C., Siligardi, G., Hussain, R., Haris, P.I., and Austen, B.M. 2001. Effect of the disulfide bridge and the C-terminal extension on the oligomerization of the amyloid peptide ABri implicated in familial British dementia. *Biochemistry* **40:** 3449–3457.

El-Agnaf, O.M.A., Gibson, G., Lee, M., Wright, A., and Austen, B.M. 2004. Properties of neurotoxic peptides related to the Bri gene. *Protein Pept. Lett.* **11:** 202–212.

Ferguson, N., Berriman, J., Petrovich, M., Sharpe, T.D., Finch, J.T., and Fersht, A.R. 2003. Rapid amyloid fibril formation from the fast-folding WW domain FBP28. *Proc. Natl. Acad. Sci.* **100:** 9814–9819.

Fernandez Escamilla, A.M., Rousseau, F., Schymkowitz, J., and Serrano, L. 2004. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat. Biotech.* **22:** 1302–1306.

Ferrara, P., Apostolakis, J., and Caflisch, A. 2002. Evaluation of a fast implicit solvent model for molecular dynamics simulations. *Proteins* **46:** 24–33.

Fersht, A.R. 1999. *Structure and mechanism in protein science*. Freeman and Co., New York.

Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R.D., and Bairoch, A. 2003. Expasy: The proteomics server for in depth protein knowledge and analysis. *Nucleic Acids Res.* **31:** 3784–3788.

Gazit, E. 2002. A possible role for π-stacking in the self-assembly of amyloid fibrils. *FASEB J.* **16:** 77–83.

Gordon, D.J., Balbach, J.J., Tycko, R., and Meredith, S.C. 2004. Increasing the amphiphilicity of an amyloidogenic peptide changes the β-sheet structure in the fibrils from antiparallel to parallel. *Biophys. J.* **86:** 428–434.

Gsponer, J., Habertuer, U., and Caflisch, A. 2003. The role of side-chain interactions in the early steps of aggregation: Molecular dynamics simulations of an amyloid-forming peptide from the yeast prion Sup35. *Proc. Natl. Acad. Sci.* **100:** 5154–5159.

Haggqvist, B., Naeslund, J., Sletten, K., Westermark, G.T., Mucchiano, G., Tjernberg, L.O., Nordstedt, C., Engstroem, U., and Westermark, P. 1999. Medin: An integral fragment of aortic smooth muscle cell-produced lactadherin forms the most common human amyloid. *Proc. Natl. Acad. Sci.* **96:** 8669–8674.

Hill, A.F., Joiner, S., Linehan, J., Desbruslais, M., Lantos, P.L., and Collinge, J. 2000. Species-barrier-independent prion replicates in apparently resistant species. *Proc. Natl. Acad. Sci.* **97:** 10248–10253.

Horwich, A.L. and Weissman, J.S. 1997. Deadly conformations-protein misfolding disease. *Cell* **89:** 499–510.

Hwang, W., Zhang, S., Kamm, R.D., and Karplus, M. 2004. Kinetic control of dimer structure formation in amyloid fibrillogenesis. *Proc. Natl. Acad. Sci.* **101:** 12916–12921.

Jaroniec, C.P., MacPhee, C.E., Astrof, N.S., Dobson, C.M., and Griffin, R.G. 2002. Molecular conformation of a peptide fragment of transthyretin in an amyloid fibril. *Proc. Natl. Acad. Sci.* **99:** 16748–16753.

Jarrett, J., Berger, E.P., and Lansbury Jr., P.T. 1993. The carboxyl terminus of the β amyloid protein critical for the seeding of amyloid formation: Implications for the pathogenesis of Alzheimer's disease. *Biochemistry* **32:** 4693–4697.

Jenkins, J. and Pickersgill, R. 2001. The architecture of parallel β-helices and related folds. *Prog. Biophys. Mol. Biol.* **77:** 111–115.

Jimenez, J.L., Nettleton, E.J., Bouchard, M., Robinson, C.V., Dobson, C.M., and Saibil, H.R. 2002. The protofilament structure of insulin amyloid fibrils. *Proc. Natl. Acad. Sci.* **99:** 9196–9201.

Jones, S., Manning, J., Kad, N.M., and Radford, S.E. 2003. Amyloid-forming peptides from b_2 microglobulin—Insights into the mechanism of fibril formation in vitro. *J. Mol. Biol.* **325:** 249–257.

Kangas, H., Paunio, T., Kalkkinen, N., Jalanko, A., and Peltonen, L. 1996. In vitro expression analysis shows that the secretory form of Gelsolin is

the sole source of amyloid in Gelsolin-related amyloidosis. *Hum. Mol. Genet*. **5:** 1237–1244.

Kayed, R., Bernhagen, J., Greenfield, N., Sweimeh, K., Brummer, H., Voelter, W., and Kapurniotu, A. 1999a. Conformational transitions of islet amyloid polypeptide (IAPP) in amyloid formation in vitro. *J. Mol. Biol*. **287:** 781–796.

———. 1999b. Partial molar volume, surface area, and hydration changes for equilibrium unfolding and formation of aggregation transition state: High-pressure and cosolute studies on recombinant human IFN-γ. *J. Mol. Biol*. **287:** 781–796.

Kelly, J. 1998. The alternative conformations of amyloidogenic proteins and their multi-step assembly pathways. *Curr. Opin. Struct. Biol*. **8:** 101–106.

King, C.Y., Tittmann, P., Gross, H., Gebert, R., Aebi, M., and Wuethrich, K. 1997. Prion-inducing domain 2–114 of yeast Sup35 protein transforms in vitro into amyloid-like filaments. *Proc. Natl. Acad. Sci*. **94:** 6618–6622.

Konno, T., Murata, K., and Nagayama, K. 1999. Amyloid-like aggregates of a plant protein: A case of sweet tasting protein, monellin. *FEBS Lett*. **454:** 122–126.

Kozin, S.A., Bertho, G., Mazur, A.K., Rabesona, H., Girault, J.P., Haerlthé, T., Takahashi, M., Debey, P., and Hui Bon Hoa, G. 2001. Sheep prion protein synthetic peptide spanning helix 1 and β-strand 2 residues 142–166 shows β-hairpin structure in solution. *J. Biol. Chem*. **49:** 46364–46370.

Kusumoto, Y., Lomakin, A., Teplow, D.B., and Benedek, G.B. 1998. Temperature dependence of amyloid β-protein fibrillization. *Proc. Natl. Acad. Sci*. **95:** 12277–12282.

Linding, R., Schymkowitz, J., Rousseau, J., Diella, F., and Serrano, L. 2004. A comparative study of the relationship between protein structure and β-aggregation in globular and intrinsically disordered proteins. *J. Mol. Biol*. **342:** 345–353.

Litvinovich, S.V., Brew, S.A., Aota, S., Akiyama, S.K., Haudenschild, C., and Ingham, K.C. 1998. Formation of amyloid like fibrils by self-association of a partially unfolded fibronectin type III module. *J. Mol. Biol*. **280:** 245–258.

Lomakin, A., Chung, D.S., Benedek, G.B., Kirschner, D.A., and Teplow, D.B. 1996. On the nucleation and growth of amyloid β-protein fibrils: Detection of nuclei and quantitation of rate constants. *Proc. Natl. Acad. Sci*. **93:** 1125–1129.

Lomakin, A., Teplow, D.B., Kirschner, D.A., and Benedek, G.B. 1997. Kinetic theory of fibrillogenesis of amyloid β-protein. *Proc. Natl. Acad. Sci*. **94:** 7942–7947.

Makin, O.S., Atkins, E., Sikorski, P., Johansson, J., and Serpell, L.C. 2005. Molecular basis for amyloid fibril formation and stability. *Proc. Natl. Acad. Sci*. **102:** 315–320.

Marcotte, E.M. and Eisenberg, D. 1999. Chicken prion tandem repeats form a stable, protease-resistant domain. *Biochemistry* **38:** 667–676.

Margittai, M. and Langen, R. 2004. Template-assisted filament growth by parallel stacking of τ. *Proc. Natl. Acad. Sci*. **101:** 10279–10283.

Massi, F. and Straub, J.E. 2001. Energy landscape theory for Alzheimer's amyloid β-peptide fibril elongation. *Proteins* **42:** 217–229.

Matthews, D. and Cooke, B. 2003. The potential for transmissible spongiform encephalopathies in non-ruminant livestock and fish. *Rev. Sci. Tech*. **22:** 283–296.

McGaughey, G.B., Gagné, M., and Rappé, A.K. 1998. π-Stacking interaction. *J. Biol. Chem*. **273:** 15458–15463.

Michelitsch, M.D. and Weissman, J.S. 2000. A census of glutamine/asparagine-rich regions: Implications for their conserved function and the prediction of novel prions. *Proc. Natl. Acad. Sci*. **97:** 11910–11915.

Munishkina, L.A., Fink, A.L., and Uversky, V.U. 2004. Conformational prerequisites for formation of amyloid fibrils from histones. *J. Mol. Biol*. **342:** 1305–1324.

Nahway, N. 1989. *The Merck index*. Merck and Co., Inc., Whitehouse Station, NJ.

Nguyen, J., Baldwin, M.A., Cohen, F.E., and Prusiner, S.B. 1995. Prion protein peptides induce α-helix to β-sheet conformational transitions. *Biochemistry* **34:** 4186–4192.

Nichols, W.C., Dwulet, F.E., Liepnieks, J., and Benson, M.D. 1988. Variant apolipoprotein AI as a major constituent of a human hereditary amyloid. *Biochem. Biophys. Res. Commun*. **156:** 762–768.

Padrick, S.B. and Miranker, A.D. 2002. Islet amyloid: Phase partitioning and secondary nucleation are central to the mechanism of fibrillogenesis. *Biochemistry* **41:** 4694–4703.

Perutz, M.F. 1999. Glutamine repeats and neurodegenerative diseases: Molecular aspects. *Trends Biochem. Sci*. **24:** 58–64.

Perutz, M.F., Johnson, T., Suzuki, M., and Finch, J.T. 1994. Glutamine repeats as polar zippers: Their possible role in inherited neurodegenerative diseases. *Proc. Natl. Acad. Sci*. **91:** 5355–5358.

Porat, Y., Stepensky, A., Ding, F.X., Naider, F., and Gazit, E. 2003. Completely different amyloidogenic potential of nearly identical peptide fragments. *Biopolymers* **69:** 161–163.

Prusiner, S.B. 1988. Prions. *Proc. Natl. Acad. Sci*. **95:** 13363–13383.

Rochet, J.C. and Lansbury Jr., P.T. 2000. Amyloid fibrillogenesis: Themes and variations. *Curr. Opin. Struct. Biol*. **10:** 60–68.

Scherzinger, E., Lurz, R., Turmaine, M., Mangiarini, L., Hollenback, B., Hasenbank, R., Bates, G.P., Davies, S.W., Lehrack, H., and Wanker, E. 1997. Huntingtin-encoded polyglutamine expansions form amyloid-like protein aggregates in vitro and in vivo. *Cell* **90:** 549–558.

Soto, C. and Castilla, J. 2004. The controversial protein-only hypothesis of prion propagation. *Nat. Med*. **10:** S63–S67.

Stefani, M. and Dobson, C.M. 2003. Protein aggregation and aggregate toxicity: New insights into protein folding, misfolding diseases and biological evolution. *J. Mol. Med*. **81:** 678–699.

Tanaka, M., Morishima, I., Akagi, T., Hashikawa, T., and Nukina, N. 2001. Intra and intermolecular β-pleated sheet formation in glutamine-repeat inserted myoglobin as a model for polyglutamine diseases. *J. Biol. Chem*. **276:** 45470–45475.

Tartaglia, G.G., Cavalli, A., Pellarin, R., and Caflisch, A. 2004. The role of aromaticity, exposed surface, and dipole moment in determining protein aggregation rates. *Protein Sci*. **13:** 1939–1941.

Tartaglia, G.G., Pellarin, R., Cavalli, A., and Caflisch, A. 2005. Organism complexity anti-correlates with proteomic β-aggregation propensity. *Protein Sci*. (this issue).

Tenidis, K., Waldner, M., Bernhagen, J., Fischle, W., Bermann, M., Weber, M., Merkle, M., Voelter, W., Brunner, H., and Kapurniotu, A. 2000. Identification of a penta- and hexapeptide of Islet amyloid polypeptide IAPP with amyloidogenic and cytotoxic properties. *J. Mol. Biol*. **295:** 1055–1071.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. Clustal W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res*. **22:** 4673–4680.

Tjernberg, L., Hosia, W., Bark, N., Thyberg, J., and Johansson, J. 2002. Charge attraction and β-propensity are necessary for amyloid fibril formation from tetrapeptides. *J. Biol. Chem*. **277:** 43243–43246.

Torok, M., Milton, S., Kayed, R., Wu, P., Intire, T.M., Glabe, C., and Langen, R. 2002. Structural and dynamic features of Alzheimer Aβ peptide in amyloid fibrils studied by site-directed spin labeling. *J. Biol. Chem*. **277:** 40810–40815.

Ueda, K., Fukushima, H., Masliah, E., Xia, Y., Iwai, A., Yoshimoto, M., Otero, D.A., Kondo, J., Ihara, Y., and Saitoh, T. 1993. Molecular cloning of cDNA encoding an unrecognized component of amyloid in Alzheimer disease. *Proc. Natl. Acad. Sci*. **90:** 11282–11286.

Vanik, D.L., Surewicz, K.A., and Surewicz, W.K. 2004. Molecular basis of barriers for intraspecies transmissibility of mammalian prions. *Mol. Cell* **14:** 139–145.

von Bergen, M., Friedhoff, P., Biernat, J., Heberle, J., Mandelkow, E.M., and Mandelkow, E. 2000. Assembly of τ protein into Alzheimer paired helical filaments depends on a local sequence motif (^{306}VQIVYK^{311}) forming β-structure. *Proc. Natl. Acad. Sci*. **97:** 5129–5134.

Weidemann, A., Konig, G., Bunke, D., Fisher, P., Salbaum, J.M., Masters, C.L., and Beyreuther, K. 1989. Identification, biogenesis and localization of precursors of Alzheimer's disease A4 amyloid protein. *Cell* **57:** 115–126.

Westermark, P., Wernstedt, C., Wilander, E., Hayden, D.W., O'Brien, T.D., and Johnson, K.H. 1987. Amyloid fibrils in human insulinoma and islets of Langerhans of the diabetic cat are derived from a neuropeptide-like protein also present in normal islet cells. *Proc. Natl. Acad. Sci*. **84:** 3881–3885.

Westermark, G.T., Engstrom, U., and Westermark, P. 1992. The N-terminal segment of protein AA determines its fibrillogenic propensity. *Biochem. Biophys. Res. Commun*. **182:** 27–32.

Williams, A.D., Portelius, E., Kheterpal, I., Guo, J.T., Cook, K.D., Xu, Y., and Wetzel, R. 2004. Mapping Aβ amyloid fibril secondary structure using scanning proline mutagenesis. *J. Mol. Biol*. **335:** 833–842.

Zoete, V., Michielin, O., and Karplus, M. 2003. Protein-ligand binding free energy estimation using molecular mechanics and continuum electrostatics. Application to HIV-1 protease inhibitors. *J. Comput. Aided Mol. Des*. **17:** 861–880.

5.3 Organism complexity anti-correlates with proteomic β-aggregation propensity. [Protein Sci. 2005, *14*, 2735]

FOR THE RECORD

Organism complexity anti-correlates with proteomic β-aggregation propensity

GIAN GAETANO TARTAGLIA,[1] RICCARDO PELLARIN,[1] ANDREA CAVALLI, AND AMEDEO CAFLISCH

Department of Biochemistry, University of Zürich, CH-8057 Zürich, Switzerland

(RECEIVED March 23, 2005; FINAL REVISION June 23, 2005; ACCEPTED June 24, 2005)

Abstract

We introduce a novel approach to estimate differences in the β-aggregation potential of eukaryotic proteomes. The approach is based on a statistical analysis of the β-aggregation propensity of polypeptide segments, which is calculated by an equation derived from first principles using the physicochemical properties of the natural amino acids. Our analysis reveals a significant decreasing trend of the overall β-aggregation tendency with increasing organism complexity and longevity. A comparison with randomized proteomes shows that natural proteomes have a higher degree of polarization in both low and high β-aggregation prone sequences. The former originates from the requirement of intrinsically disordered proteins, whereas the latter originates from the necessity of proteins with a stable folded structure.

Keywords: aggregation; protein aggregation propensity; proteome; intrinsically disordered proteins

Supplemental material: see www.proteinscience.org

Even proteins not implicated in amyloid diseases have been shown to form fibrils in vitro under denaturing conditions, indicating that fibrillogenesis is a common feature of polypeptide chains, which can form intermolecular backbone–backbone hydrogen bonds (Chiti et al. 1999, 2003) and favorable side-chain interactions (Azriel and Gazit 2001; Gsponer et al. 2003; Makin et al. 2005). Although in lower eukaryotes amyloid fibrils could represent an inheritable phenotype related to specific cellular functions (Osherovich and Weissman 2002; Osherovich et al. 2004; Si et al. 2003b), the cytotoxicity of prefibrillar aggregates (Bucciantini et al. 2002) and their association with diseases such as Alzheimer's, Parkinson's, Huntington's, prion disease, cystic fibrosis, and type II diabetes (Kelly 1998; Rochet and Lansbury 2000) suggest that amyloid aggregates are generally dangerous for higher eukaryotes (Dobson 1999; Stefani and Dobson 2003).

We have previously developed an equation to predict the propensity for ordered aggregation, which solely requires the polypeptide sequence as input (Tartaglia et al. 2004, 2005). Our model is based on the physicochemical properties of the residues and takes into account both amino acid composition and positional information. The aggregation propensity π_{il} of an l-residue segment starting at position i in the sequence is evaluated as

$$\pi_{il} = \phi_{il}\ \Phi_{il} \quad (1)$$

The factor Φ_{il} contains exponential functions and is position-dependent

$$\Phi_{il} = e^{A_{il}+B_{il}+C_{il}} \quad (2)$$

where A_{il}, B_{il}, and C_{il} are functionals related to the aromaticity, β-propensity, and charge, respectively. The fac-

[1]These authors contributed equally to this work.

Reprint requests to: Gian Gaetano Tartaglia or Amedeo Caflisch, Department of Biochemistry, University of Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland; e-mail: gian@bioc.unizh.ch or caflisch@bioc.unizh.ch; fax: +41-44-635-68-62.

Article published online ahead of print. Article and publication date are at http://www.proteinscience.org/cgi/doi/10.1110/ps.051473805.

tor ϕ_{il} depends almost exclusively on the amino acid composition

$$\phi_{il} = \left[\prod_{j=i}^{i+l-1} \left(\frac{S_j^a}{\hat{S}^a} \theta^{\uparrow\uparrow} + \frac{S_j^p}{\hat{S}^p} \theta^{\uparrow\downarrow} \right) \frac{\hat{S}^t}{S_j^t} \frac{\hat{\sigma}}{\sigma_j} \right]^{1/l} \quad (3)$$

where S_i^a, S_i^p, S_i^t, and σ_i—weighted by their average over the 20 standard amino acids (hatted values)—are the side-chain apolar, polar, total water-accessible surface area, and solubility, respectively. The functionals $\theta^{\uparrow\uparrow}$ and $\theta^{\uparrow\downarrow}$ include positional effects and reflect the parallel or anti-parallel tendency to aggregate if the majority of residues is apolar or polar, respectively. Details of the method are presented in the preceding paper (Tartaglia et al. 2005).

In the present work, we analyze complete proteomes of several eukaryotes to identify changes of β-aggregation propensity through organisms of different complexity. The 32,869 entries belonging to the human proteome database (Supplemental Material, Table 1) were decomposed in stretches of different sizes (5, 50, and 150 residues) to compute the β-aggregation propensity with Equation 1 and build the normalized histogram of β-aggregation propensity distribution, APD (Fig. 1A). For each stretch size, the distribution is found to be nonsymmetric with respect to the average and skewed to the left, indicating that there are more stretches with low β-aggregation propensity (left tail of APD) than with high propensity (right tail). As pointed out in our previous study, short stretches are preferable to long stretches for the analysis of β-aggregation propensity because the latter contain folding features that deteriorate the signal-to-noise ratio (Tartaglia et al. 2005).

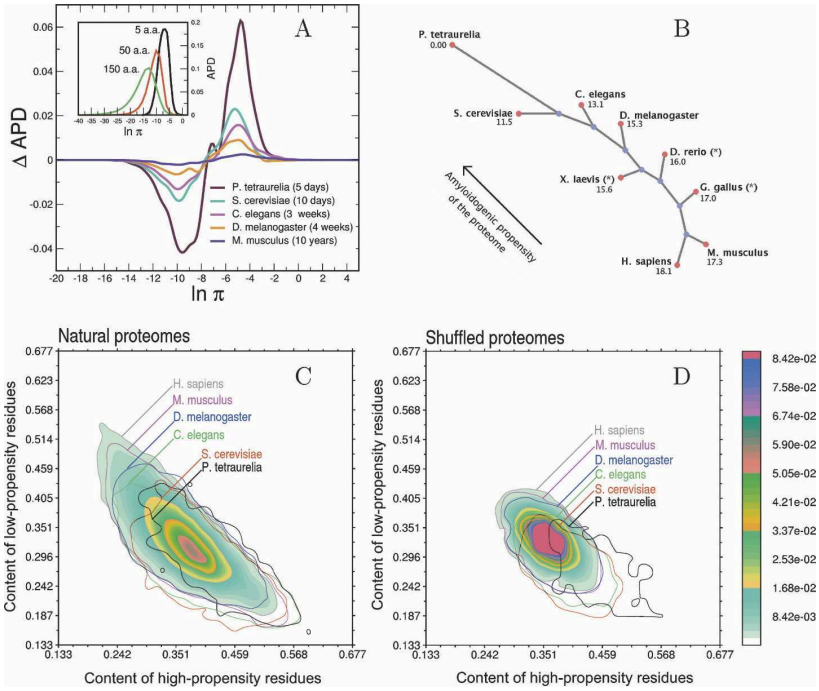


Figure 1. (*A*) (*Inset*) Distribution of the number of human polypeptide sequences as a function of β-aggregation propensity (APD) at three different window sizes. (*Main plot*) APD differences with respect to *H. sapiens* for complete proteomes of *M. musculus*, *D. melanogaster*, *C. elegans*, *S. cerevisiae*, and *P. tetraurelia* (window size of five residues). Life spans of organisms are reported in parentheses. (*B*) Unrooted tree diagram derived from the APD deviation (Equation 4). The deviation is computed from *P. tetraurelia* as a reference and magnified by a factor of 1000. The arrow indicates that lower eukaryotes have more high-propensity and fewer low-propensity stretches. This diagram is built using Phylodraw with the Fitch and Margoliash (1967) clustering algorithm. Data labeled with * belong to incomplete proteomes. (Phylodraw is available at http://pearl.cs.pusan.ac.kr/phylodraw/.) (*C*) Normalized histogram of the number of proteins as a function of the content of residues enriched in low-propensity and high-propensity stretches. Global contours are shown for all proteomes by solid lines. Isofrequency regions are shown for the human proteome, where red color indicates the most populated area, while blue fading color indicates the least-populated areas. (*D*) Same as *C* for shuffled proteomes.

Hence, a window size of five residues was used to analyze complete proteomes of *Homo sapiens*, *Mus musculus*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, and *Paramecium tetraurelia* (Supplemental Material, Table 1). Nonhuman eukaryotes show a larger amount of high-propensity stretches and a smaller amount of low-propensity stretches compared with *H. sapiens* (Fig. 1A). Moreover, a clear trend is found with the increasing complexity of the organisms and their lifetime. To quantify this trend it is useful to introduce the APD deviation between two proteomes, x and y

$$d_{xy} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(APD_x(\pi_i) - APD_y(\pi_i)\right)^2} \qquad (4)$$

where the β-aggregation propensity π is calculated by Equation 1 (Tartaglia et al. 2005) and i runs over the total number of bins N ($N = 100$) in the APD histogram. With the addition of the proteomes of *Danio rerio*, *Xenopus laevis*, and *Gallus gallus*, the APD deviation was used to build the tree diagram of Figure 1B. Except for the inversion between the amphibious *X. laevis* and the fish *D. rerio* (whose proteomes are not complete), the tree of Figure 1B is similar to the phylogenetic tree of cytochrome c (Dayhoff et al. 1972). Thus, the deviation calculated from *P. tetraurelia*, d_{xP}, is an observable able to rank proteomes of organisms of increasing complexity. It is interesting to compare the amino acid frequencies in APD tails—defined for a subtended area of 0.05 in the histogram of Figure 1A—with amino acid frequencies in entire proteomes (Table 1). This analysis reveals that for all proteomes stretches with low β-aggregation propensity are rich in *A*, *G*, *H*, *K*, *P* and *R*, whereas high-propensity stretches in *C*, *F*, *I*, *L*, *N*, *Q*, *V*, and *Y*. Figure 1C is a two-dimensional histogram that shows the number of proteins as a function of the content of residues enriched in low-propensity stretches and the content of residues predominant in high-propensity stretches. By increasing the organism complexity, the number of proteins with low-propensity residues increases, while the number of proteins with high-propensity residues decreases. A comparison with randomized proteomes is useful to further investigate the significance of such trends. Randomized proteomes were generated by shuffling amino acids within complete proteomes and keeping unchanged the global amino acid composition, number, and length of proteins. We stress that the β-aggregation propensity of five-residue stretches cannot differentiate natural and shuffled proteomes, because short segments describe mainly effects of the amino acid composition. Yet, differences between natural and shuffled proteomes are enhanced when residues belonging to low-/high-propensity stretches are used for the analysis of entire proteins. Comparing Figure 1, C and D, it is evident that shuffled proteomes are less spread. In other words, natural proteomes reveal a sensible increase of sequences with residues predominant in low-propensity stretches as well as residues enriched in high-propensity stretches. While the amino acid global composition of proteomes is almost identical in higher eukaryotes, the content of low-propensity stretches increases significantly, indicating a clear change of protein features from proteome to proteome.

Table 1. *Amino acid frequencies in left or right* APD *tails of* H. sapiens *divided by their corresponding frequency in the whole proteome*

	A	C	D	E	F	G	H	I	K	L
Left/total	**1.1**	0.2	0.4	0.5	0.4	**1.3**	**1.6**	0.5	**2.1**	0.5
Right/total	0.7	**2.4**	0.8	0.7	**2.7**	0.6	0.5	**1.6**	0.3	**1.8**
	M	N	P	Q	R	S	T	V	W	Y
Left/total	0.4	0.2	**3.3**	0.3	**2.8**	0.6	0.5	0.7	0.4	0.1
Right/total	0.8	**1.5**	0.2	**1.2**	0.2	0.7	0.8	**1.2**	0.8	**2.7**

Values exceeding 1.0 are shown in bold. Similar frequencies were found for all the proteomes.

It has recently been shown that natively unfolded proteins (or intrinsically disordered proteins, IDPs) are implicated in cellular regulation, signaling, and assembly/disassembly of macromolecular complexes (Dunker et al. 2002; Ward et al. 2004; Oldfield et al. 2005). The absence of a fixed structure suggests functional implications, which are required in complex organisms (Koonin et al. 2002). Interestingly, a larger diffusion of IDPs is found in higher eukaryotes than in lower eukaryotes and prokaryotes (Dunker et al. 2002; Liu et al. 2002; Linding et al. 2004). Using data from X-ray crystallography, nuclear magnetic resonance, and circular dichroism, Williams et al. (2001) found a high percentage of *P*, *R*, *K*, *G*, *A*, *Q*, *S*, and *E* in nonfolded segments of proteins, and *F*, *Y*, *C*, *L*, *V*, *N*, and *W* in folded segments. Except for *Q*, *S*, and *E*, Williams' finding is in agreement with our tail composition analysis (Table 1), indicating that residues enriched in aggregating stretches promote both folding and β-aggregation, whereas residues predominant in stretches with low β-aggregation propensity are also enriched in IDPs.

To better understand the relationship between β-aggregation propensity and protein structure, we analyzed the APDs of polypeptide segments that assume a regular secondary structure, as well as IDPs (Supplemental Material, Table 1). As shown in Figure 2A, strands have more β-aggregation potential than helices,

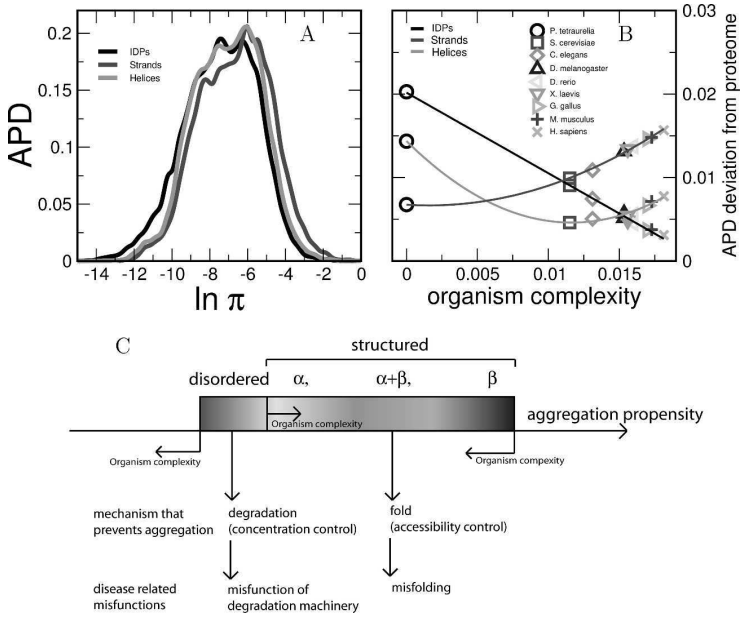


Figure 2. (*A*) APDs of five-residue stretches belonging to intrinsically disordered proteins (IDPs) or regular secondary structure elements within folded proteins and IDPs. (*B*) Deviation between the APD of entire proteomes and the APD of segments belonging to regular secondary structure or IDPs as a function of the organism complexity. The organism complexity is measured by the APD deviation from *P. tetraurelia*, d_{xP}. Solid lines are drawn solely to guide the eye. (*C*) From lower to higher eukaryotes, the decrease of β-aggregation propensity is related to the increase of intrinsically disordered proteins.

and IDPs are the least prone to aggregate, in agreement with Linding's analysis (Linding et al. 2004). Moreover, from lower to higher eukaryotes the APD deviation with respect to IDP decreases, while the APD deviation from strands increases (Fig. 2B,C). The APD deviation of helices does not follow a monotonic trend and slowly increases from *S. cerevisiae* to *H. sapiens*. Compared with strands, helices display a lower amount of aggregation stretches, but it has to be mentioned that the transition helix-strand generates amyloidogenesis in some proteins (Selkoe 1996; Prusiner 1997).

To quantify interspecies shifts of amino acid compositions in the APD tails, we fitted the amino acid frequencies as a linear function of the APD deviation from *P. tetraurelia*, d_{xP} (see Equation 4)

$$f_x^a = shift^a \; d_{xP} + cst^a \tag{5}$$

where f_x^a is the frequency of the amino acid a in the proteome x, $shift^a$ is the slope of the fit, and cst^a is the intercept. The sign "+" or "−" of the $shift^a$ was interpreted as a measure for the depletion or the enrichment of the amino acid a from *P. tetraurelia* to *H. sapiens*. Shifts obtained from high-confidence fits (Pearson's correlation >0.80; Supplemental Material, Table 2) are

- Right tails, i.e., high propensity: Decrease of *Q*, *N*, *Y*, and *K* and increase of *L*, *V*, *A*, *W*, *R*, *H*, *G*, and *P*.
- Left tails, i.e., low propensity: Decrease of *K*, *I*, *F*, and *N* and increase of *P*, *A*, *G*, *R*, *S*, and *E*.

Interestingly, the decrease of *Q*, *N*, and *Y* in the right tails was already observed in higher eukaryote prion homologs of the yeast Sup35 prion protein (Balbirnie et al. 2001; Si et al. 2003a; Theis et al. 2003) and suggests that the trend does not affect only a specific family of proteins. In addition, we speculate that the increase of *L*, *V*, *A*, and *W* in the right tail is a consequence of the optimization of the "hydrophobic core" to stabilize the native state (Kellis et al. 1989; Richards and Lim 1993; Dill et al. 1995; Stefani and Dobson 2003).

The functional role of aggregation phenotypes in multicellular eukaryotes is still a matter of debate. Recently, it has been observed that the neuronal protein CPEB of *Aplysia californica* behaves like a prion switch that regulates long-term synaptic changes asso-

ciated with memory storage (Si et al. 2003a,b). The switch mechanism involves the aggregation of the CPEB N terminus, rich in *Q*- and *N*- repeats that are missing in mammalian isoforms of CPEB (Theis et al. 2003). Motivated by these observations, we analyzed the data set of proteins expressed in neurons (Supplemental Material , Table 1). For a given proteome, the neuronal APD perfectly overlaps with the APD of the total proteome (data not shown), indicating that neuronal proteins are a descriptive subset of the total proteome and do not follow any specific trend. We thus cannot draw conclusions on particular links between memory mechanisms and aggregation phenotypes.

It has been shown that the frequency of *N* and *Q* repeats does not represent an observable able to describe amyloidogenic trends of proteomes (Michelitsch and Weissman 2000; Osherovich and Weissman 2002). Our findings indicate that to quantify aggregation trends, it is crucial to use an observable, such as the β-aggregation propensity, which accounts for the aggregation contribution of all amino acids including positional information.

In conclusion, we have introduced a novel approach to compare proteomes, which is based on the statistical analysis of ordered-aggregation propensity. From *P. tetraurelia* to *H. sapiens*, we have shown that proteomes of higher and more long-lived eukaryotes contain fewer sequences with high β-aggregation propensity and are accrued in proteins with low β-aggregation propensity. We also observed that, compared with random proteomes, natural proteomes are enriched in proteins with low β-aggregation potential, as well as proteins with high β-aggregation potential. Such polarization is a consequence of the dual evolutive requirement of IDPs with low β-aggregation propensity, as well as proteins with a stable fold, which comes at the cost of higher β-aggregation propensity. In the future, we plan to use gene ontology annotations of proteins with high predicted β-aggregation propensity to obtain insights into the specific role of some of the amyloidogenic proteins of unknown function.

Electronic supplemental material

This section contains two tables: Table 1 contains information for databases used in the article (origin of data sets, number of entries of the databases, and number of stretches used in our analysis); Table 2 contains fitting parameters for the amino acid shifts (see Equation 5).

Acknowledgments

We thank Dr. A.G. Abebe and M. Cecchini for very interesting discussions. This work was supported by the Swiss National Science Foundation and the NCCR "Neural Plasticity and Repair."

References

Azriel, R. and Gazit, E. 2001. Analysis of the minimal amyloid-forming fragment of the islet amyloid polypeptide. *J. Biol. Chem.* **276:** 34156–34161.

Balbirnie, M., Grothe, R., and Eisenberg, D. 2001. An amyloid-forming peptide from the yeast prion Sup35 reveals a dehydrated β-sheet structure for amyloid. *Proc. Natl. Acad. Sci.* **98:** 2375–2380.

Bucciantini, M., Giannoni, E., Chiti, F., Baroni, F., Formigli, L., Zurdo, J., Taddei, N., Ramponi, G., Dobson, C.M., and Stefani, M. 2002. Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases. *Nature* **416:** 507–511.

Chiti, F., Calamai, M., Taddei, N., Stefani, M., Ramponi, G., and Dobson, C.M. 1999. Studies of the aggregation of mutant proteins in vitro provide insights into the genetics of amyloid diseases. *Proc. Natl. Acad. Sci.* **99:** 16419–16426.

Chiti, F., Stefani, M., Taddei, N., Ramponi, G., and Dobson, C.M. 2003. Rationalization of the effects of mutations on peptide and protein aggregation rates. *Nature* **424:** 805–808.

Dayhoff, M.O., Park, C.M., and McLaughlin, P.J. 1972. *Atlas of protein sequence and structure*. National Biomedical Research Foundation, Silver Spring, MD.

Dill, K.A., Yue, K., Fiebig, K.M., Yee, D.P., Thomas, P.D., and Chan, H.S. 1995. Principles of protein folding—A perspective from simple exact models. *Protein Sci.* **4:** 561–602.

Dobson, C.M. 1999. Protein misfolding, evolution and disease. *Trends Biochem. Sci.* **24:** 329–332.

Dunker, A.K., Brown, C.J., Lawson, J.D., Iakoucheva, L.M., and Obradovic, Z. 2002. Intrinsic disorder and protein function. *Biochemistry* **41:** 6574–6582.

Fitch, W.M. and Margoliash, E. 1967. Construction of phylogenetic tree. *Science* **155:** 279–284.

Gsponer, J., Habertuer, U., and Caflisch, A. 2003. The role of side-chain interactions in the early steps of aggregation: Molecular dynamics simulations of an amyloid-forming peptide from the yeast prion Sup35. *Proc. Natl. Acad. Sci.* **100:** 5154–5159.

Kellis, J.T., Nyberg, K., and Fersht, A.R. 1989. Energetics of complementary side-chain packing in a protein hydrophobic core. *Biochemistry* **28:** 4914–4922.

Kelly, J.W. 1998. The alternative conformations of amyloidogenic proteins and their multi-step assembly pathways. *Curr. Opin. Struct. Biol.* **8:** 101–106.

Koonin, E.V., Wolf, Y.I., and Karev, G.P. 2002. The structure of the protein universe and genome evolution. *Nature* **420:** 218–223.

Linding, R., Schymkowitz, J., Rousseau, J., Diella, F., and Serrano, L. 2004. A comparative study of the relationship between protein structure and β-aggregation in globular and intrinsically disordered proteins. *J. Mol. Biol.* **342:** 345–353.

Liu, J., Tau, H., and Rost, B. 2002. Loopy proteins appear conserved in evolution. *J. Mol. Biol.* **322:** 53–64.

Makin, O.S., Atkins, E., Sikorski, P., Johansson, J., and Serpell, L.C. 2005. Molecular basis for amyloid fibril formation and stability. *Proc. Natl. Acad. Sci.* **102:** 315–320.

Michelitsch, M.D. and Weissman, J.S. 2000. A census of glutamine/asparagine-rich regions: Implications for their conserved function and the prediction of novel prions. *Proc. Natl. Acad. Sci.* **97:** 11910–11915.

Oldfield, C.L., Cheng, Y., Cortese, M.S., Brown, C.J., Uversky, V.N., and Dunker, A.K. 2005. Comparing and combining predictors of mostly disordered proteins. *Biochemistry* **44:** 1989–2000.

Osherovich, L.Z. and Weissman, J.S. 2002. The utility of prions. *Dev. Cell* **2:** 143–151.

Osherovich, L.Z., Cox, B.S., Tuite, M.F., and Weissman, J.S. 2004. Dissection and design of yeast proteins. *PLoS Biol.* **2:** 442–451.

Prusiner, S.B. 1997. Prion diseases and the BSE crisis. *Science* **278:** 245–251.

Richards, F.M. and Lim, W. 1993. An analysis of packing in the protein folding problem. *Q. Rev. Biophys.* **26:** 423–498.

Rochet, J.C. and Lansbury Jr., P.T. 2000. Amyloid fibrillogenesis: Themes and variations. *Curr. Opin. Struct. Biol.* **10:** 60–68.

Selkoe, D.J. 1996. Amyloid β-protein and the genetics of Alzheimer's disease. *J. Biol. Chem.* **271:** 18295–18298.

Si, K., Giustetto, M., Etkin, A., Hsu, R., Janisiewicz, A.M., Miniaci, M.C., Kim, J.H., Zhu, H., and Kandel, E.R. 2003a. A neuronal isoform of CPEB regulates local protein synthesis and stabilizes synapse-specific long-term facilitation in aplysia. *Cell* **115:** 893–904.

Si, K., Linquist, S., and Kandel, E.R. 2003b. A neuronal isoform of the aplysia CPEB has prion-like properties. *Cell* **115:** 879–891.

Stefani, M. and Dobson, C.M. 2003. Protein aggregation and aggregate toxicity: New insights into protein folding, misfolding diseases and biological evolution. *J. Mol. Med.* **81:** 678–699.

Tartaglia, G.G., Cavalli, A., Pellarin, R., and Caflisch, A. 2004. The role of aromaticity, exposed surface, and dipole moment in determining protein aggregation rates. *Protein Sci.* **13:** 1939–1941.

———. 2005. Prediction of aggregation rate and aggregation-prone segments in polypeptide sequences. *Protein Sci.* (this issue).

Theis, M., Si, K., and Kandel, E.R. 2003. Two previously undescribed members of the mouse CPEB family of genes and their inducible expression in the principal cell layers of the hippocampus. *Proc. Natl. Acad. Sci* **100:** 9602–9607.

Ward, J.J., Sodhi, J.S., McGuffin, L.J., Buxton, B.F., and Jones, D.T. 2004. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.* **337:** 635–645.

Williams, R.M., Obradovic, Z., Mathura, V., Braun, W., Garner, E.C., Young, J., Takayama, S., Brown, C.J., and Dunker, A.K. 2001. The protein non-folding problem: Amino acid determinans of intrinsic order and disorder. *Pac. Symp. Biocomput.* **200:** 89–100.

6 Simplified model for simulations of amyloid aggregation

The ability of proteins to fold efficiently to their native conformation may protect them from side-reactions such as amorphous and amyloid aggregation: evolution have shaped the folding and the aggregation pathways in such a way that the native functional state is *in vivo* a thermodynamically metastable structure [67]. One of the key questions currently unanswered is at which point the folding and aggregation pathways meet.

A detailed picture of oligomer formation and the influence of the protein landscape on the kinetics and the thermodynamics of the process is still under debate. The important issues concern the understanding of the multistep process, including the early stage of the oligomerization process and the protofibril to fibril conversion. As mentioned earlier (see figure 1), amyloid fibrillization of dispersed polypeptides belongs to the more general theory of nucleated growth polymerization. According to this mechanism the rate limiting step for the amyloid formation is the nucleation phase, where nuclei are generated. This phase is followed by the elongation phase, where nuclei indefinitely grow by polypeptide addition, and they eventually assume the shape of structured fibrils. Prior to nucleation, during the lag phase, a variety of soluble oligomers can be explored, and depending on the system, they are observed to promote or deplete the nucleation and the elongation phase. The nucleus is the least stable species on the aggregation pathway, and it is in pre-equilibrium with the monomers. According to this theory the nucleus, once formed, can either degrade to monomers and to non-amyloid oligomers, or progress to fibril.

Interestingly, disease associated mutants of α-synuclein are related to the acceleration of oligomerization rather than fibrillization [68]. Also in the case of β-amyloid peptide, the ”arctic” mutation, linked to an early onset form of Alzheimer disease, has been observed to enhance the protofibril, while keeping constant the fibril elongation kinetics *in vitro* [69], but also displayed high amyloidogenicity *in vivo* [70]. Many recent works have also highlighted the importance of protofibrillar assemblies, mainly addressing the question whether protofibrils are on or off pathway intermediates. Hydrogen exchange

measurements revealed that Abeta40 protofibrils have indeed a β-sheet rich protected structure, revealing a possible on-pathway placement [71]. Interestingly apomyoglobin protofibrils arise from a multi-step reaction by a random nucleation mechanism in which the polypeptide chains sample many different aggregated conformations [72].

In section 6.1 a coarse grained model of an amphipathic polypeptide is described. The conformational landscape of the molecule has been schematized such that only two states are considered: the amyloid-competent (β) and the amyloid-protected (π) states (see figure 3). The latter represents the ensemble of all polypeptide conformations that are not compatible with the fibril arrangement. The folding reaction expressed by equation (1) can be adapted for an isolated polypeptide chain that undergoes to a reversible isomerization from the conformation π to the conformation β:

$$\pi \underset{k_{\beta\pi}}{\overset{k_{\pi\beta}}{\rightleftharpoons}} \beta \tag{8}$$

The free energy difference $\Delta G_{\beta\pi}$ between the two states is expressed by the equilibrium constant $K_{\beta\pi}$, that is the the ratio between the populations of the states π and β

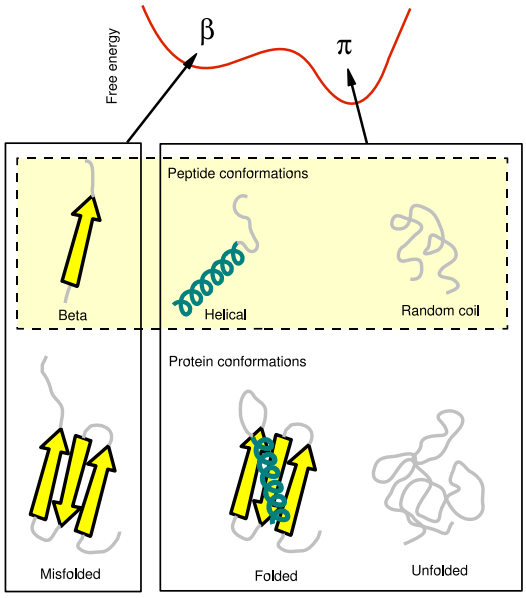


Figure 3: Mapping of real polypeptides conformations onto the free energy landscape of the coarse grained model. Top, in the yellow box, peptide conformations, bottom, protein conformations.

of the isolated monomer. The $\Delta G_{\beta\pi}$ is thus a measure of the intrinsic accessibility of the β state for the polypeptide in its monomeric form, and can be interpreted as a amyloid propensity. The variation of this observable, operated by the force field parameter $dE = E_\pi - E_\beta$, influences the kinetics and the thermodynamics of fibrillization, leading to a wide and differentiated scenario. The range of behaviors investigated can be roughly classified as β-stable model, which corresponds to a polypeptide with the β state that is thermodynamically accessible, and a β-unstable model, that represents a polypeptide for which the β state is only marginally accessible.

Depending on the parameter dE, the fibrillization occurs through alternative nucleation steps. Decreasing the population of the amyloid state β, the kinetics of fibril elongation and nucleation are dramatically slowed down. β-unstable models (i.e. dE <2.0) display a lag time that is 1 or 2 orders of magnitude greater than that obtained by β-stable model (dE=0). This can be explained by the fact that the size of the critical nuclei is increasing together with the population of the π state. Furthermore the kinetic stability of oligomers that appears at the lag phase, increases with decreasing dE.

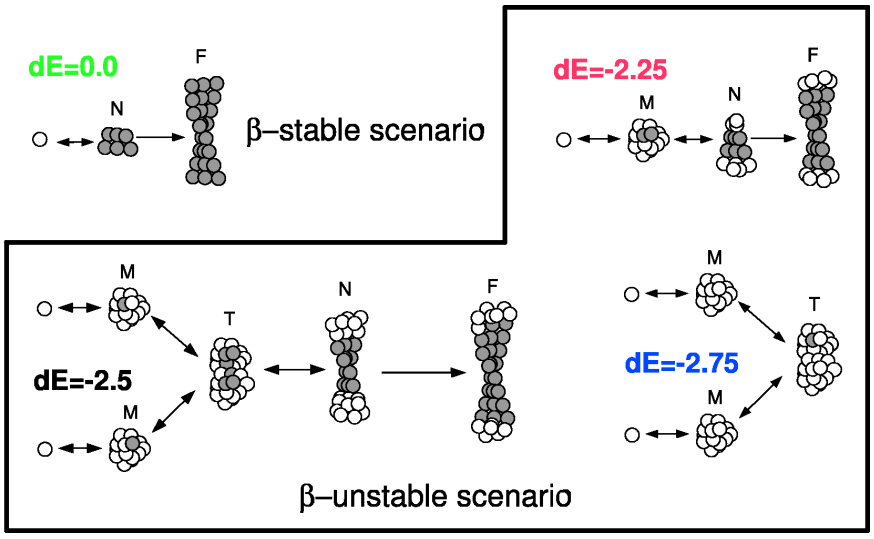


Figure 4: Observed nucleation pathways. Dark and white circles represents molecules in the amyloid-competent (β) and amyloid protected (π) states. β-stable potentials (dE=0) nucleate without intermediates, while β-unstable can nucleate either through micelle-sized oligomers (dE=-2.25) or oligomers larger than micelle (dE=-2.5). A further stabilization of the protected state prevent fibril formation (dE=-2.75). Legend: (M) micelle, (N) nucleus, (T) transient oligomer, (F) fibril.

Behaviors described in this work can help interpreting the variety of experimental observations, and derived rational models for the nucleation mechanism. Remarkably, a predicted direct correlation between the lag time and the elongation time has been subsequently confirmed by experimental measurements of the kinetics of aggregation for different amyloid polypeptides [73], a dependence that could be explained considering that the mechanisms behind nucleation and elongation are similar. Further details of the coarse-grained model, such as the parameterization, the dependence of the fibril formation kinetics upon changing the barrier between the π and the β states, and the seeded fibril growth are discussed in section 6.2.

The parameter dE also affects the pathway of fibril formation and the stability of the intermediates (see section 6.3). The decrease of the β-aggregation propensity results in multiple elongation pathways, with intermediates consisting of protofibrils that are smaller and less structured than the final fibril. When the β state is strongly unfavored, the templated formation of an additional protofilament on lateral surface is a collective transition. The simulation results of this work might help to interpret the polymorphism in the intermediates of amyloid production [47], the dependence of the pathways upon changing the external conditions [74], the growth of fibril assisted by lateral scaffold [75], the production of protofibrils upon mutation [69] and the structural models of protofibrils [71].

6.1 Interpreting the aggregation kinetics of amyloid peptides. [J. Mol. Biol. 2006, *360*, 882]

doi:10.1016/j.jmb.2006.05.033 *J. Mol. Biol.* (2006) **360**, 882–892

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Interpreting the Aggregation Kinetics of Amyloid Peptides

Riccardo Pellarin and Amedeo Caflisch*

Department of Biochemistry
University of Zürich
Winterthurerstrasse 190
CH-8057 Zürich
Switzerland

Amyloid fibrils are insoluble mainly β-sheet aggregates of proteins or peptides. The multi-step process of amyloid aggregation is one of the major research topics in structural biology and biophysics because of its relevance in protein misfolding diseases like Alzheimer's, Parkinson's, Creutzfeld-Jacob's, and type II diabetes. Yet, the detailed mechanism of oligomer formation and the influence of protein stability on the aggregation kinetics are still matters of debate. Here a coarse-grained model of an amphipathic polypeptide, characterized by a free energy profile with distinct amyloid-competent (i.e. β-prone) and amyloid-protected states, is used to investigate the kinetics of aggregation and the pathways of fibril formation. The simulation results suggest that by simply increasing the relative stability of the β-prone state of the polypeptide, disordered aggregation changes into fibrillogenesis with the presence of oligomeric on-pathway intermediates, and finally without intermediates in the case of a very stable β-prone state. The minimal-size aggregate able to form a fibril is generated by collisions of oligomers or monomers for polypeptides with unstable or stable β-prone state, respectively. The simulation results provide a basis for understanding the wide range of amyloid-aggregation mechanisms observed in peptides and proteins. Moreover, they allow us to interpret at a molecular level the much faster kinetics of assembly of a recently discovered functional amyloid with respect to the very slow pathological aggregation.

Keywords: amyloid fibril; aggregation intermediate; fibril formation kinetics; multiple pathways; Alzheimer's disease

*Corresponding author

Introduction

Amyloid fibrils are polypeptide aggregates with a core structure consisting of β-sheets whose strands are perpendicular to the fibril axis, and the backbone hydrogen bonds are parallel to it.[1] Despite the medical relevance of several diseases linked to amyloidosis, important questions about the formation kinetics of the early ordered aggregates remain unanswered. The initial phase of fibril formation is of particular interest because experimental evidence has accumulated indicating that soluble oligomeric precursors, rather than the fibrils, might be the toxic species.[2–5] However, the transient nature of oligomeric precursors makes it difficult to shed light on their formation process or structure. Additional interest in a better understanding of aggregation kinetics is being spurred by the very recent discovery of functional amyloids in mammalian cells,[6] which is challenging the view that amyloid is always cytotoxic. It is likely that aggregation kinetics have to be very fast in functional amyloids in contrast to the very slow progress of pathological protein aggregation.

Theoretical models have been developed to investigate the amyloid aggregation mechanism[7–9] and predict the rates[10] but important assumptions like the irreversible association of polypeptide chains onto the fibril[7,10] are not consistent with experimental evidence.[11,12] Computer simulations using low-resolution models, which employ a simplified representation of protein geometry and energetics, have provided insights into the basics physical principles underlying protein aggregation in general[13–15] and ordered amyloid aggregation.[16–20] Yet, they do not explain the wide range of aggregation scenarios emerging from a variety of biophysical measurements.[21,22] Computer simulations at the atomic level of detail[23–29] have shed some light on

E-mail address of the corresponding author:
caflisch@bioc.unizh.ch

oligomeric aggregates and the very early steps of fibril formation. However, all-atom simulations of the kinetics of fibril formation are beyond what can be done with modern computers.

To overcome such computational limitations, we approximate a polypeptide by a coarse-grained model consisting of ten beads (Figure 1) and simulate 125 monomers in a cubic box. The monomer has internal (dihedral) flexibility and a free energy profile with two minima at the amyloid-competent state β and the amyloid-protected state π (Figure 1(d)). The latter state represents the ensemble of all polypeptide conformations that are not compatible with the cross β arrangement in a fibril, e.g. α-helical or random coil structures. An important result obtained with the coarse-grained model is that the kinetic phases and fibril formation mechanism depend on the choice of a single parameter, the relative stability of π and β states ($dE = E_\pi - E_\beta$). Notably, very different aggregation scenarios are observed by varying dE from the β-unstable ($dE \leq -2$ kcal/mol) to the β-stable ($dE \geq 0$ kcal/mol) model. The simulation results are used to interpret biophysical data on several peptides and proteins.

Results

Most of the results were obtained at a temperature of 310 K and a concentration of 8.5 mM unless specified otherwise.

Multi-step process

The range of aggregation kinetics is shown in Figures 2 and 3 while two illustrative snapshots from a simulation of the β-unstable polypeptide model ($dE = -2.5$ kcal/mol) are given in Figure 4. Three different kinetic phases are evident: lag, elongation (or growth), and final monomer-fibril equilibrium. The variable length of the lag phase and the higher heterogeneity at longer lag times are indicative of a stochastic nucleation.[30] Fibril formation is much slower for the β-unstable model than the β-stable model ($dE = 0$ kcal/mol). The model parameter dE also affects the elongation kinetics; β-unstable models display a slower elongation rate (Figure 2(b)). Interestingly, the significant anti-correlation between the length of the lag phase and the elongation rate for different values of dE is consistent with the kinetic analysis of single-point mutants of Aβ40 (compare Figure 2(a) and (b) with Figures 1 and 2 of Christopeit *et al.*,[31] respectively). The distribution function $p(N)$ of the oligomer size N evaluated at the lag phase or the final equilibrium is depicted in Figure 3. Within a given phase the peaks of the $p(N)$ distribution can be interpreted as stable oligomeric species. The monomer peak ranges from $N = 1$ to 7, the micellar peak from $N = 8$ to 60, and the fibril peak from $N = 61$ to 125. The micellar peak is present for the $dE = -2.5$ kcal/mol model at the lag phase, but disappears at the final equilibrium, where the fibril and the monomers are the only co-existing species. The height of the peaks depends on the relative stability of the β-competent state as well as the total monomer

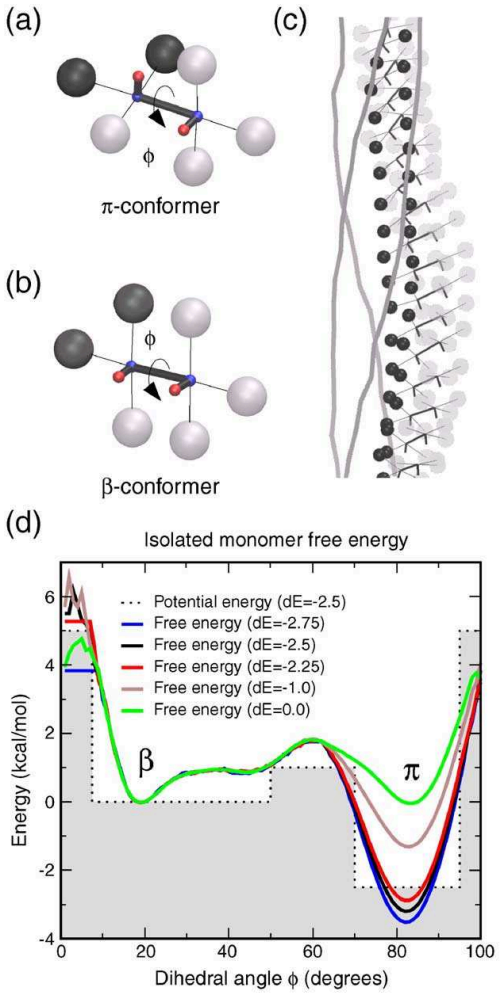


Figure 1. The model. Sticks and beads representation of the monomer in the amyloid-protected state π (a) and the amyloid-competent state β (b). The large spheres are hydrophobic (black) and hydrophilic (gray). The small red and blue spheres indicate the negative and positive partial charges that make up the "backbone" dipoles. The circular arrow indicates the rotatable bond ϕ. (c) Structure of a filament embedded in a four-filament fibril. Hydrophobic spheres are in the core of the fibril, while hydrophilic spheres are on the outside. All monomers are in the β state. The three other filaments are shown by gray ribbons. Each filament is twisted and the four filaments are intertwined. (d) The dotted line is the dihedral potential with a $dE = -2.5$ kcal/mol energy difference between amyloid-protected and amyloid-competent state. The five continuous lines represent the free energy profile of the isolated monomer for five different dihedral potentials.

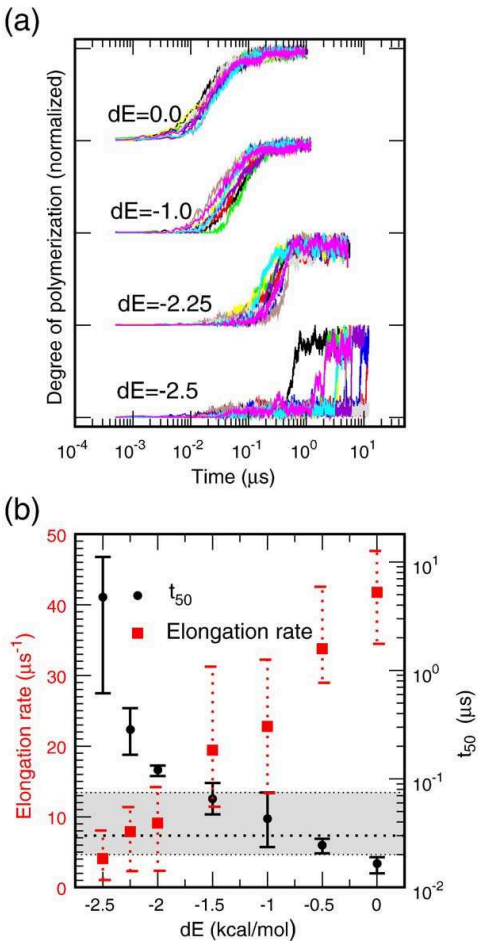


Figure 2. Effect of relative stability of the amyloid-competent state on nucleation and elongation kinetics. (a) Time series of the fraction of ordered aggregation evaluated at four values of the protected state stability dE. Ten independent simulations are shown for each dE value. The degree of polymerization is normalized to the maximum for each curve. Note that the average value at the plateau is about 10% smaller for the $dE=-2.5$ kcal/mol than the $dE=0.0$ kcal/mol model. It is not possible to directly compare the slopes of the curves (i.e. elongation rates) at different values of dE because of the logarithmic scale of the x-axis. (b) Influence of dE on the kinetics of the system. The time needed to reach 50% of the maximal amplitude t_{50} (black circles and y-axis legend on the right) and the elongation rate (red squares and y-axis legend on the left) are displayed for seven dE values. Symbols represent the average value of ten independent runs and the error bars are the maximum and minimum values. The broken line and the gray band indicate the average and the max–min values for the time of micelle formation, respectively. All simulations were performed at a temperature of 310 K and a concentration of 8.5 mM.

concentration. For the β-stable potential $dE=0.0$ kcal/mol the micelle peak is not observed at any concentration value. With increasing concentration the monomer and micelle peak distributions are skewed towards high N values because multi-monomer collisions and multi-micellar collisions, respectively, transiently generate oligomers of a larger size.

Micellar aggregates

Micelle-like non-fibrillar aggregates are observed only for the models with an unstable amyloid-competent state ($dE \leq -2$ kcal/mol). Furthermore, the comparison between the lag times with the time of micelle formation (gray area in Figure 2(b)) shows that the fibril formation kinetics of the β-unstable and β-stable models are slower and faster than micelle formation, respectively. In fact, micelles are intermediates consisting mainly of monomers in the π state (Figure 4(a)) whereas the polymerization of β-stable monomers directly yields fibrils. The number of monomers in the micelle is 18±5 at the concentration value of 8.5 mM and 22±8 at 62 mM, and the critical concentration of micelle formation is 4.36 mM (see Figure 5). The micellar size, its weak dependence on the concentration (see Figure 5), and the on-pathway location are in good agreement with recent fluorescence quenching data obtained for the Alzheimer's Aβ40 peptide.[32]

Mechanism of nucleation

The nucleation properties of the system are investigated by evaluating the probability of fibril formation for β-subdomains, i.e. the clusters of interacting β-monomers. The nucleus, defined as the oligomer containing a β-subdomain with a 50% probability to form a fibril, shows an increasing size upon destabilization of the β-state. The number of monomers in the nucleus is about four for the $dE=0.0$ kcal/mol model, while it is 27 and 40 for the β-unstable models $dE=-2.25$ kcal/mol and $dE=-2.5$ kcal/mol, respectively (Figure 6(a) and (b)). Significantly different nucleation mechanisms are observed upon reduction of the relative stability of the amyloid-competent state (Figure 6(c)). In the β-stable model ($dE=0$ kcal/mol), the nucleus size is sub-micellar and nucleation is simply the aggregation of monomers in the β-state. For the β-unstable models, nucleus formation requires either spatial proximity within a micelle of several monomers in the β-state ($dE=-2.25$ kcal/mol) or collision of two micelles with merging of their β-subdomains ($dE=-2.5$ kcal/mol). The lack of lag phase in the simulation of a "seeding experiment" (with the $dE=-2.5$ kcal/mol model, see Supplementary Data) is consistent with the nucleation dependence of the aggregation process. For the highly β-unstable model ($dE=-2.75$ kcal/mol) amyloid fibril formation is inhibited by formation of non-fibrillar aggregates having a very low density of monomers in the amyloid-competent state (Figure 6(b) and (c)).

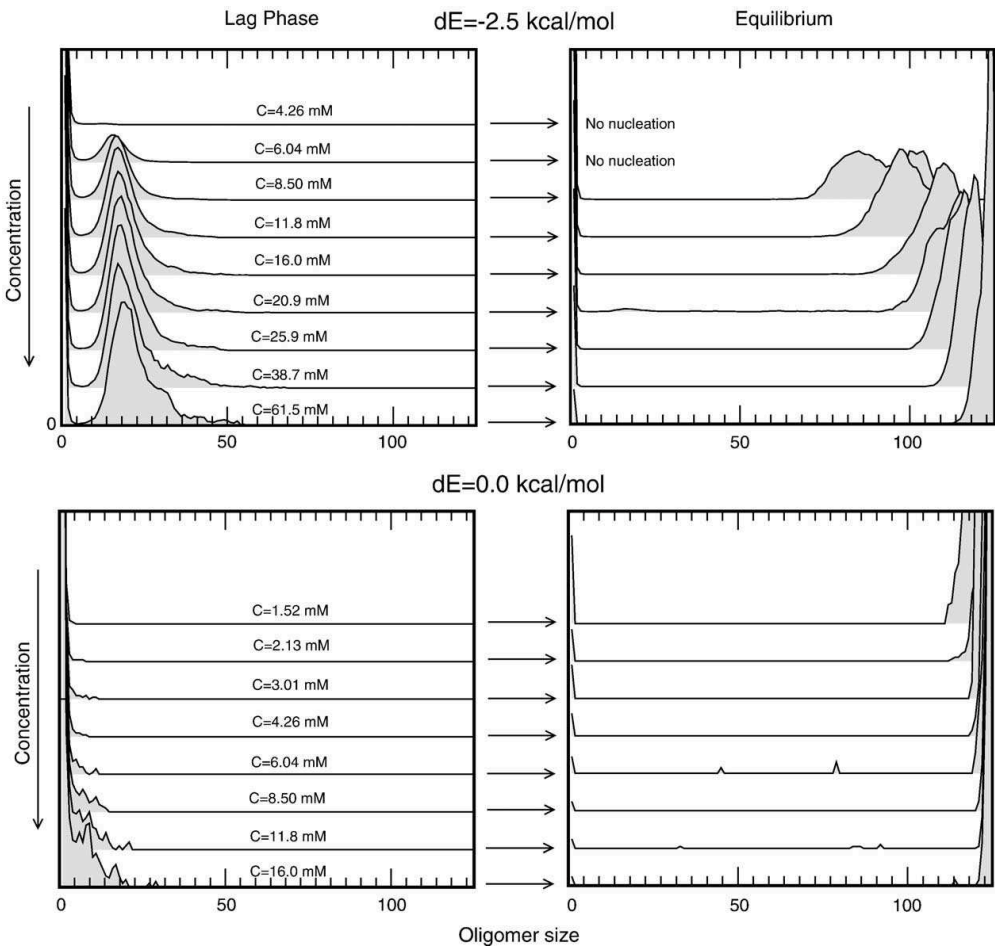


Figure 3. Oligomer size histograms of the $dE = -2.5$ kcal/mol potential (top) and $dE = 0.0$ kcal/mol potential (bottom) calculated in the lag phase (left) and the final equilibrium (right). Lag phase and equilibrium histograms evaluated at the same value of the concentration are reported in the same row. The z-dimension represents the relative probability.

Concentration effects

Another major difference between the models with unstable and stable amyloid-competent state is the dependence on the total monomer concentration of the nucleation and elongation kinetics. In agreement with the above-mentioned mechanism of nucleation, the β-unstable model nucleates only at concentration values larger than the critical concentration of micelle formation, whereas the β-stable model nucleates even at lower concentrations (Figure 7(a)). Furthermore, during the lag phase the micelle concentration increases linearly with the total monomer concentration for the β-unstable potential (see Figure 5) in agreement with neutron and light scattering data of Aβ40 during the lag phase,[33] while the concentration of dispersed monomers remains constant and equal to the critical concentration of micelle formation. Hence, the concentration dependence of the lag phase time for the β-unstable potential indicates that micelles promote the nucleation. Interestingly, the dependence of the rate of elongation on the concentration decreases significantly by increasing the stability of the protected state π (Figure 7(b)). The reduced concentration dependence originates from competitive polymerizations, i.e. the elongation of the fibril and the presence of micelles. This observation for the β-unstable model is a consequence of the monomer–micelle equilibrium, which maintains a nearly constant concentration of isolated monomers.[7]

Final monomer–fibril equilibrium

The final part of the simulations with $dE = -2.25$ or -2.5 kcal/mol is characterized by a dynamic equilibrium between monomers and fibril[11] because the micellar aggregates disappear after the elongation

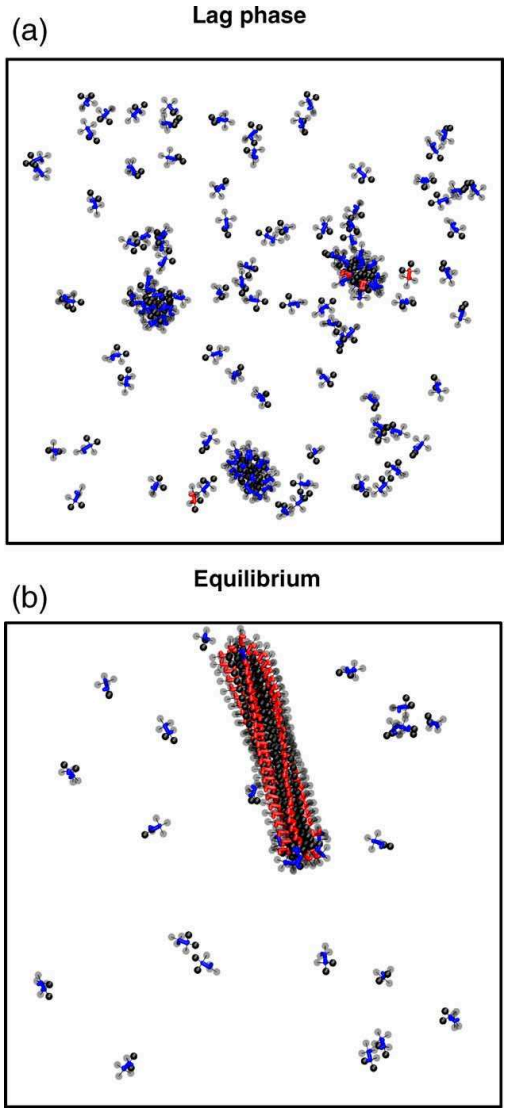


Figure 4. Kinetic phases for the β-unstable model. Snapshots from the lag phase (a) and the final monomer–fibril equilibrium (b) for a simulation with $dE = -2.5$ kcal/mol and 8.5 mM concentration. Hydrophobic and hydrophilic spheres are black and gray, respectively. The backbone is blue for π-state monomers and red for β-state monomers to emphasize that micelles consist mainly of the former (a) whereas the fibril is made up of the latter (b).

phase (Figures 3 and 4). Fibrils consist of bundles of four intertwined filaments (Figures 1(c) and 4(b)). Their caps are disordered and host monomers in the π state whose population correlates with the relative stability of the π *versus* β state. The final equilibrium observed in the coarse-grained model simulations indicates that the assumption of irreversible association of polypeptide chains onto the fibril[10] is not justified. Recently, a molecular recycling mechanism has been observed by a combination of NMR spectroscopy and mass spectroscopy for an amyloid fibril formed from an SH3 domain.[12] To evaluate the recycling time of the coarse-grained model, simulations of mature fibrils in equilibrium with dispersed monomers are analyzed for the β-unstable potential ($dE = -2.5$ kcal/mol) at different concentration values by monitoring the number of the unrecycled monomers $N_u(t)$ (Figure 8). In two of nine simulations, $N_u(t)$ goes to zero within 4 μs showing that all monomers initially belonging to the fibril have been recycled. The time for recycling half of the monomers incorporated into a fibril is of the order of 2–20 μs and is independent of the total monomer concentration (see inset of Figure 8). Such "molecular recycling" indicates that, despite their ordered arrangement, fibrils are dynamic assemblies.

Discussion

The detailed description of the kinetics and thermodynamics of the coarse-grained model suggests some general conclusions concerning the aggregation mechanism of amyloidogenic peptides and proteins. The striking variety of fibril formation mechanisms mainly depends on the relative stability of the amyloid-competent state of the monomer. Despite the essentially identical structure of the final fibril, ordered aggregation of the β-stable model follows a downhill pathway[34,35] without intermediates,

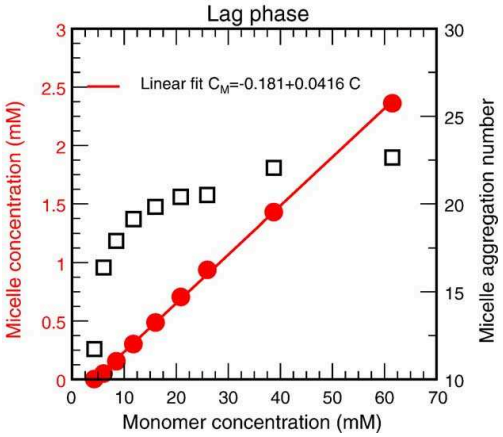


Figure 5. Lag phase of the $dE = -2.5$ kcal/mol potential. The micelle concentration C_M (red circles, *y*-axis, legend on the left) shows a linear increase when plotted as a function of the total concentration of monomers, *C*. The red continuous line is a linear fit whose parameters are reported in the graph. The *x*-axis intercept of the red continuous line is the critical concentration of micelle formation (4.36 mM). The micelle aggregation number (squares, *y*-axis, legend on the right) reaches a plateau at monomer concentration larger than 20 mM.

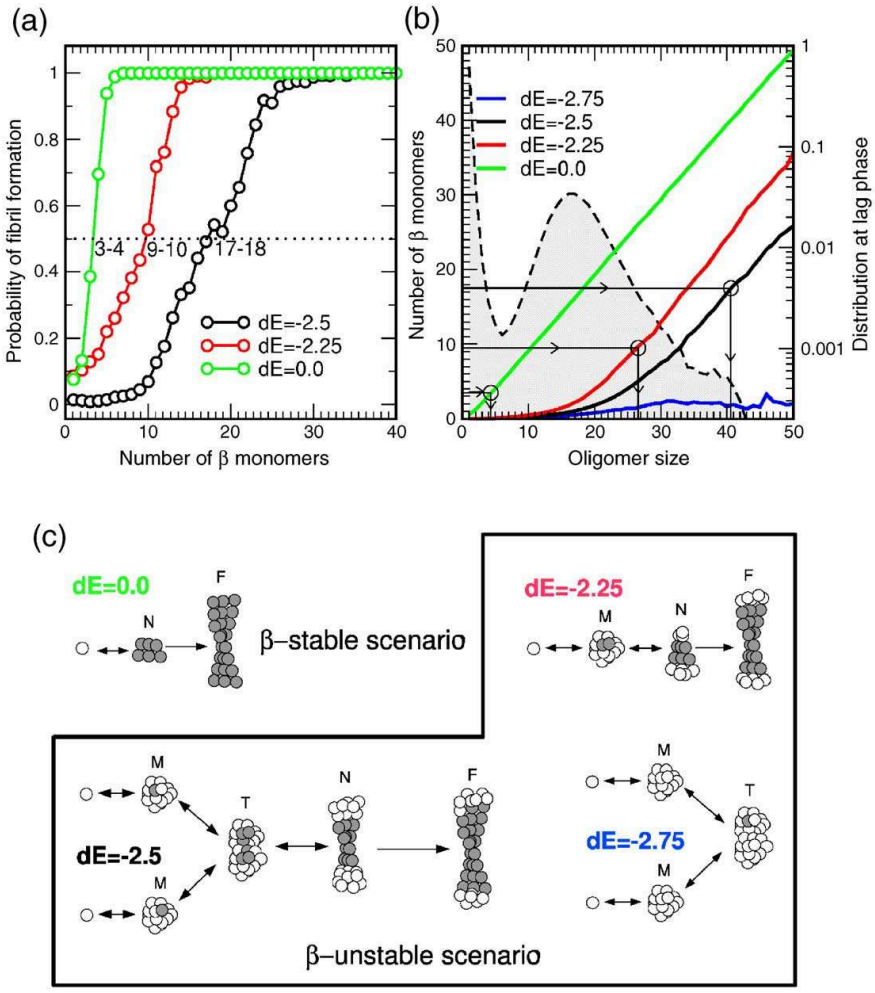


Figure 6. (a) Probability of fibril formation as a function of the size of the β-subdomain evaluated for three different dE values at a concentration of 8.5 mM. The integer ranges close to the dotted line are the β-subdomain sizes at probability equal to 0.5 (see the text for nucleus definition). (b) Average number of β-state monomers as a function of the oligomer size. The β-subdomain sizes of the left plot are reported in ordinate and the corresponding oligomer sizes are indicated by arrows. As an example, for $dE = -2.25$ kcal/mol, the nucleus consists of a β-subdomain of size 9–10 incorporated in a micelle of about 27 monomers. The broken line is the oligomer size distribution calculated at the lag phase for $dE = -2.5$ kcal/mol and concentration of 8.5 mM (see Figure 3). (c) Observed nucleation scenarios. Black and white circles represent the amyloid-competent conformer β and amyloid-protected conformer π, respectively. β-Stable monomers nucleate without intermediates, while β-unstable monomers can nucleate either through micelle-sized oligomers ($dE = -2.25$) or transient oligomers larger than micelle ($dE = -2.5$). A further stabilization of the protected state prevents fibril formation ($dE = -2.75$). M, micelle; N, nucleus; T, transient oligomer; F, fibril.

while fibrillization of the β-unstable model occurs with the presence of on-pathway oligomers. In other words, high and low β-prone sequences show completely different kinetic behaviors. This simulation result provides a basis to understand the more than four orders of magnitude faster kinetic assembly of a functional amyloid with respect to Aβ under identical experimental conditions.[6]

An unstructured peptide with a marginally stable β-prone state (e.g. Aβ40[32,36] or the islet amyloid polypeptide[37]) as well as the α/β protein acylphosphatase[22] visit oligomeric systems in the lag phase, and have a very weak dependence of the elongation rate on concentration due to the monomer–micelle equilibrium. This mechanism corresponds to the nucleated conformational conversion proposed by Serio *et al.*[38] Other examples of the β-unstable model include phosphoglycerate kinase (PGK), an α/β protein, which was investigated by light scattering, circular dichroism and electron

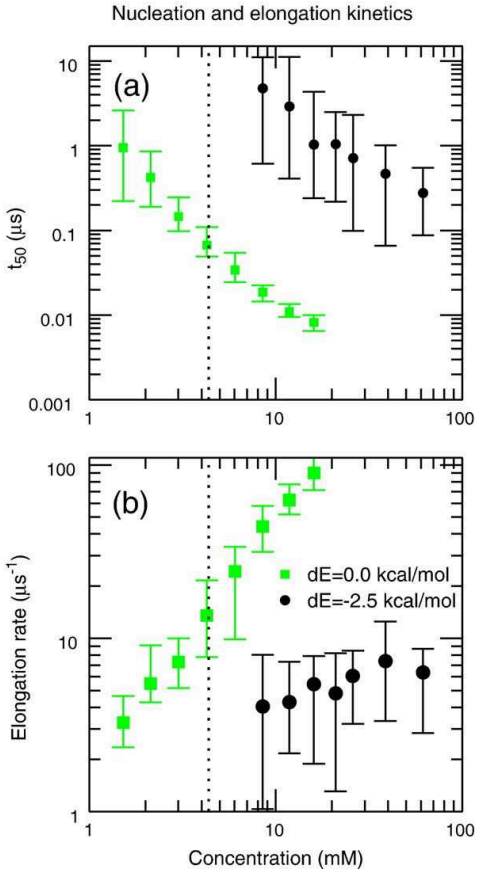


Figure 7. (a) Effect of concentration on the lag phase time t_{50} for β-unstable $dE=-2.5$ kcal/mol (black circles) and β-stable $dE=0.0$ kcal/mol (green squares) models. The symbols represent the average value calculated on 15 simulations for $dE=-2.5$ and ten simulations for $dE=0.0$. The error bars represent the minimum and the maximum value. (b) Effect of concentration on the elongation rate of the fibril. Symbols and error bars are as for (a). The vertical dotted line indicates the critical concentration of micelle formation.

microscopy. The amyloid formation of PGK was shown to be a two-stage process where critical oligomers with low β content assembly to form protofibril and fibrils of increasing cross-β structure.[9] A similar mechanism was proposed by Xu *et al.* for the full length yeast prion-like protein Sup35 using atomic force microscopy.[39] Both aforementioned mechanisms are described by the β-unstable model, which is observed to nucleate through an isolated micelle or transiently associated micelles (Figure 4(c)). On the other hand, a functional and non-pathological amyloid in mammalians[6] as well as oligopeptides with amyloid-prone sequence (e.g. the heptapeptide GNNQQNY from the N-terminal domain of the Sup35,[40,41] or the blocked diphenylalanine[42]) lack on-pathway intermediates and correspond to the β-stable model. Most importantly, it is not necessarily the stability of the folded structure, as suggested recently,[35,43] to determine the aggregation mechanism but rather the relative stability of the β-prone state.[44] In fact, the simulation results can be used to interpret the different kinetic behaviors even within proteins with stable native fold. Transthyretin[35] and β_2-microglobulin[21] have a β-sheet-rich folded state and show downhill polymerization (depending on the solution conditions) because of the intrinsic β-propensity of their sequences, as observed for the β-stable model. On the contrary, mainly α-helical structures like myoglobin and the prion protein require extreme environmental conditions (high temperature for apomyoglobin)[45] and show a significant lag phase as observed for the β-unstable models. Furthermore, the importance of the relative stability of the β-prone state is consistent with the experimental observation that the changes in nucleation and elongation kinetics upon single point mutations correlate more with the β-propensity and the hydrophobicity[46,47] than the α-helical stability as recently reported for Aβ40.[31] However, the possibility to change solely the intrinsic conformational landscape of a monomer without affecting the inter-monomer non-covalent interactions represent an advantage of the coarse-grained model and simulation approach with respect to experimental methods such as mutagenesis and solvent-induced conformational changes, by which it is not possible to separate the two effects. For instance, mutation of the Aβ40 peptide, even a single point substitution, can have an influence on both the energy landscape

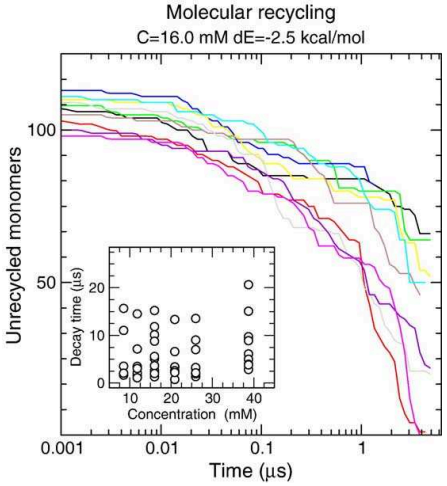


Figure 8. Number of unrecycled monomers N_u as a function of time for nine simulations started from a preformed equilibrated fibril. Simulations were performed at total concentration $C=16.0$ mM and for the $dE=-2.5$ kcal/mol model. The values of the decay time τ (i.e. the time to reach $e^{-1}N_u(0)$) at different concentrations are reported in the inset.

of the monomer and the intermolecular forces. In conclusion, a slight modification of the free energy profile is sufficient to observe a wide range of different fibril formation mechanisms providing a unifying description of the available experimental results. This finding has significant implications for the design of drugs against amyloid diseases as well as for the production of nanofibrillar materials.[48]

Materials and Methods

Model

A monomer consists of ten spherical beads arranged to have an overall amphipathic character. The relative stability of the amyloid-competent (β) and amyloid-protected (π) state is modulated by varying the dihedral energy difference of the single rotatable bond of the monomer. Interactions between monomers depend on van der Waals and electrostatic forces (Figure 1). The former approximate both steric and hydrophobic effects while the latter are dipole–dipole interactions responsible for the ordered stacking of monomers incorporated in the fibril. Details of the model are given in the Supplementary Data.

Simulation protocol

Disordered aggregation and/or fibril formation is simulated at physiological temperature (310 K) by following the time evolution (Langevin dynamics) of 125 monomers enclosed in a cubic box, at constant volume conditions, and starting from a monodisperse solution. Periodic boundary conditions are applied to avoid finite size effects. The total monomer concentration of 8.5 mM is about two orders of magnitude higher than the reported experimental concentrations of 30 μM −300 μM. The very high peptide concentration is used to increase the probability of intermolecular interactions, making the oligomerization process fast enough to be observed in a simulation time scale of about 10 μs, which requires about 25 days of CPU time. All simulations were performed with CHARMM.[49]

Dihedral free energy profile

The dihedral free energy profile reported in Figure 1(d) is evaluated for an isolated monomer as follows:

$$\Delta G(\phi) = -kT \log\left(\frac{n(\phi)}{n(\phi = 20°)}\right)$$

where $n(\phi)$ is the probability to assume a dihedral angle equal to ϕ. The β-state minimum at $\phi = 20°$ is taken as reference.

Oligomer size distribution

A clustering algorithm is used to calculate the size of the oligomeric species along the simulations. It is based on the matrix of contacts D_{ij} between labeled monomers. Given a single frame of the simulation, D_{ij} is equal to one if any sphere of monomer i is closer than 6.0 Å to any sphere of monomer j, otherwise it is zero. D_{ij} is equivalent to the first neighbor matrix, $d_{ij}^{(1)}$. The second neighbor matrix $d^{(2)}$ is constructed from $d^{(1)}$ including the neighbors of the first neighbors. The converged contact matrix $d_{ij}^{(\infty)}$ is defined by the following recursive sequence:

$$d_{ij}^{(1)} = D_{ij}$$

$$d_{ij}^{(n)} = \begin{cases} 1 & \text{if} \quad d_{ij}^{(n-1)} = 1 \\ 1 & \text{if} \quad d_{ik}^{(n-1)} = 1 \text{ and } d_{kj}^{(n-1)} = 1 \\ 0 & \text{otherwise} \end{cases}$$

$$d_{ij}^{(\infty)} = \lim_{n \to \infty} d_{ij}^{(n)}$$

which yields a block matrix of ones and zeroes. Each block represents a cluster of monomers and contains first neighbors, second neighbors, third and so on. This procedure is equivalent to a hierarchical clustering performed with a spanning tree technique.[50] With the converged contact matrix one can identify clusters (i.e. tagging each oligomer with an identification number), list monomers belonging to a specific oligomer and make statistics on the size of oligomers. The above definition of oligomeric species allows the statistical analysis of cluster size. The probability that a monomer is aggregated in a cluster of size N is:

$$p(N) = \left\langle \frac{1}{N_T} \sum_{i=1,N_T} \delta_{i,t}(N) \right\rangle_t$$

where N_T is the total number of simulated monomers, $\delta_{i,t}(N)$ is equal to 1 if the monomer i at time t is embedded in a cluster of size N, and the angular brackets are the time average. This function (termed oligomer size distribution) can be evaluated for the lag phase or the final monomer–fibril equilibrium. The elongation phase cannot be analyzed by $p(N)$ being an out of equilibrium dynamic process.

Critical concentration of micelle formation

The probability distribution $p(N)$ can be used to evaluate the critical concentration of micelle formation C^r_M. The micelle aggregation number in the lag phase is defined as:

$$N_M = \sum_{N=8}^{60} N p^{lp}(N)$$

where p^{lp} is the function $p(N)$ evaluated only in the lag phase (where there is co-existence of micelles and monomers without fibrils). The number of micelles per simulation box is $N_T p_M N_M^{-1}$, where the total number of simulated monomers N_T is 125 and p_M is the probability of a monomer being in a micelle. The micelle concentration C_M is derived from the number of micelles in the simulation volume. The micelle aggregation number N_M and concentration are plotted in Figure 5 as a function of the total monomer concentration C for the potential $dE = -2.5$ kcal/mol in the lag phase. By extrapolating a linear fit of the concentration of micelles the critical concentration of micelle formation C^r_M can be evaluated. The value is $C^r_M = 4.36$ mM.

Kinetics

The time series displayed in Figure 2(a) are evaluated by counting the number of parallel polar contacts n_p, i.e. the number of inter-monomer dipole–dipole interactions normalized to the maximum value. This observable is

approximately zero at the lag phase and increases with the degree of fibril polymerization after the nucleation. The elongation rate (Figures 2(b) and 7(b)) is evaluated by fitting the n_p time series with an exponential function, whereas t_{50} (Figures 2(b) and 7(a)) is the time needed to reach 50% of the maximum n_p amplitude. Time series of the number of inter-monomer hydrophobic contacts n_h are used to calculate the time of micelle formation (gray band in Figure 2(b)), which is the time needed to reach 50% of the amplitude at the lag phase starting from initially dissolved monomers ($n_h = 0$). See Supplementary Data for additional information.

Probability of fibril formation

β-Subdomains, defined as clusters of interacting β-monomers, are used to evaluate the probability of fibril formation shown in Figure 6(a). Given a β-subdomain A_{N_β} of size N_β, the set of pathways P_i is defined as the set of β-subdomain trajectories produced starting from A_{N_β}. The probability of fibril formation for a β-subdomain of size N_β is defined as:

$$p_{Ff}(N_\beta) = \frac{1}{M}\sum_{P_i} F(P_i)$$

where M is a normalization term that accounts for the occurrence of oligomer A_{N_β} in the simulations, and $F(P_i)$ is equal to 1 if the pathway produces a fiber, and otherwise it is 0. The number of β-monomers in an oligomer of size N, reported in Figure 6(b), is calculated from the free energy difference between the β and π states in an oligomer of size N, $\Delta G_{\beta\pi}(N)$:

$$N_\beta(N) = \frac{Ne^{\frac{\Delta G_{\beta\pi}(N)}{kT}}}{1 + e^{\frac{\Delta G_{\beta\pi}(N)}{kT}}}$$

See Supplementary Data for additional information.

Number of unrecycled monomers

The number of unrecycled monomers $N_u(t)$ is defined as follows. First, all monomers belonging to the fibril at time $t = 0$ are labeled and counted. Then, at all times $t > 0$ the monomers that never detached from the fibril are counted and the resulting number is $N_u(t)$. The number of unrecycled monomers $N_u(t)$ can be fitted with an exponential function:

$$N_u(t) = N_u(0)e^{-t/\tau}$$

where $N_u(0)$. is the initial value of monomers belonging to the fibril and τ is the decay time.

Acknowledgements

We thank Dr A. Cavalli, Dr M. Cecchini, Dr E. Guarnera and Dr G. Tartaglia for interesting discussions, and Dr S. Bernèche, Professor R. Melki and Professor B. Schuler for useful comments on the manuscript. The calculations were performed on Matterhorn, a Beowulf Linux cluster at the Informatikdienste of the University of Zurich, and we thank C. Bolliger, Dr T. Steenbock, and Dr A. Godknecht for installing and maintaining the Linux cluster. This work was supported by the Swiss National Competence Center in Research (NCCR) on Neural Plasticity and Repair.

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2006.05.033

References

1. Sunde, M. & Blake, C. (1997). The structure of amyloid fibrils by electron microscopy and X-ray diffraction. *Adv. Protein Chem.* **50**, 123–159.
2. Conway, K. A., Lee, S. J., Rochet, J. C., Ding, T. T., Williamson, R. E. & Lansbury, P. T. J. (2000). Acceleration of oligomerization, not fibrillization, is a shared property of both alpha-synuclein mutations linked to early-onset Parkinson's disease: implications for pathogenesis and therapy. *Proc. Natl Acad. Sci. USA*, **97**, 571–576.
3. Bucciantini, M., Giannoni, E., Chiti, F., Baroni, F., Formigli, L., Zurdo, J. *et al.* (2002). Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases. *Nature*, **416**, 507–511.
4. Kayed, R., Head, E., Thompson, J. L., McIntire, T. M., Milton, S. C., Cotman, C. W. & Glabe, C. G. (2003). Common structure of soluble amyloid oligomers implies common mechanism of pathogenesis. *Science*, **300**, 486–489.
5. Cleary, J. P., Walsh, D. M., Hofmeister, J. J., Shankar, G. M., Kuskowski, M. A., Selkoe, D. J. & Ashe, K. H. (2005). Natural oligomers of the amyloid-beta protein specifically disrupt cognitive function. *Nature Neurosci.* **8**, 79–84.
6. Fowler, D. M., Koulov, A. V., Alory-Jost, C., Marks, M. S., Balch, W. E. & Kelly, J. W. (2006). Functional amyloid formation within mammalian tissue. *PLoS Biol.* **4**, e6.
7. Lomakin, A., Teplow, D. B., Kirschner, D. A. & Benedek, G. (1997). Kinetic theory of fibrillogenesis of amyloid β-protein. *Proc. Natl Acad. Sci. USA*, **94**, 7942–7947.
8. Massi, F. & Straub, J. E. (2001). Energy landscape theory for Alzheimer's amyloid β-peptide fibril elongation. *Proteins: Struct. Funct. Bioinformat.* **42**, 217–229.
9. Modler, A. J., Gast, K., Lutsch, G. & Damaschun, G. (2003). Assembly of amyloid protofibrils *via* critical oligomers–a novel pathway of amyloid formation. *J. Mol. Biol.* **325**, 135–148.
10. Hall, D., Hirota, N. & Dobson, C. M. (2005). A toy model for predicting the rate of amyloid formation from unfolded protein. *J. Mol. Biol.* **351**, 195–205.
11. O'Nuallain, B., Shivaprasad, S., Kheterpal, I. & Wetzel, R. (2005). Thermodynamics of Aβ(1–40) amyloid fibril elongation. *Biochemistry*, **44**, 12709–12718.
12. Carulla, N., Caddy, G. L., Hall, D. R., Zurdo, J., Gairi, M., Feliz, M. *et al.* (2005). Molecular recycling within amyloid fibrils. *Nature*, **436**, 554–558.
13. Broglia, R. A., Tiana, G., Pasquali, S., Roman, H. E. & Vigezzi, E. (1998). Folding and aggregation of

designed proteins. *Proc. Natl Acad. Sci. USA*, **95**, 12930–12933.
14. Gupta, P., Hall, C. K. & Voegler, A. C. (1998). Effect of denaturant and protein concentrations upon protein refolding and aggregation: a simple lattice model. *Protein Sci.* **7**, 2642–2652.
15. Harrison, P. M., Chan, H. S., Prusiner, S. B. & Cohen, F. E. (1999). Thermodynamics of model prions and its implications for the problem of prion protein folding. *J. Mol. Biol.* **286**, 593–606.
16. Urbanc, B., Cruz, L., Yun, S., Buldyrev, S. V., Bitan, G., Teplow, D. B. & Stanley, H. E. (2004). *In silico* study of amyloid beta-protein folding and oligomerization. *Proc. Natl Acad. Sci. USA*, **101**, 17345–17350.
17. Nguyen, H. D. & Hall, C. K. (2004). Molecular dynamics simulations of spontaneous fibril formation by random-coil peptides. *Proc. Natl Acad. Sci. USA*, **101**, 16180–16185.
18. Jang, H., Hall, C. K. & Zhou, Y. (2004). Assembly and kinetic folding pathways of a tetrameric beta-sheet complex: molecular dynamics simulations on simplified off-lattice protein models. *Biophys. J.* **86** (1 Pt 1), 31–49.
19. Dima, R. I. & Thirumalai, D. (2002). Exploring protein aggregation and self-propagation using lattice models: phase diagram and kinetics. *Protein Sci.* **11** (5), 1036–1049.
20. Malolepsza, E., Boniecki, M., Kolinski, A. & Piela, L. (2005). Theoretical model of prion propagation: a misfolded protein induces misfolding. *Proc. Natl Acad. Sci. USA*, **102**, 7835–7840.
21. Gosal, W. S., Morten, I. J., W., H. E., Smith, D. A., Thomson, N. H. & Radford, S. E. (2005). Competing pathways determine fibril morphology in the self-assembly of β_2-microglobulin into amyloid. *J. Mol. Biol.* **351**, 850–864.
22. Plakoutsi, G., Bemporad, F., Calamai, M., Taddei, N., Dobson, C. M. & Chiti, F. (2005). Evidence for a mechanism of amyloid formation involving molecular reorganisation within native-like precursor aggregates. *J. Mol. Biol.* **351**, 910–922.
23. Ma, B. & Nussinov, R. (2002). Stabilities and conformations of Alzheimer's β-amyloid peptide oligomers ($A\beta_{16-22}$, $A\beta_{16-35}$, and $A\beta_{10-35}$): sequence effects. *Proc. Natl Acad. Sci. USA*, **99**, 14126–14131.
24. Gsponer, J., Haberthür, U. & Caflisch, A. (2003). The role of side-chain interactions in the early steps of aggregation: molecular dynamics simulations of an amyloid-forming peptide from the yeast prion Sup35. *Proc. Natl Acad. Sci. USA*, **100**, 5154–5159.
25. Klimov, D. & Thirumalai, D. (2003). Dissecting the assembly of A amyloid peptides into antiparallel β sheets. *Structure*, **11**, 295–307.
26. Wei, G., Mousseau, N. & Derreumaux, P. (2004). Sampling the self-assembly pathways of KFFE hexamers. *Biophys. J.* **87**, 3648–3656.
27. Hwang, W., Zhang, S., Kamm, R. D. & Karplus, M. (2004). Kinetic control of dimer structure formation in amyloid fibrillogenesis. *Proc. Natl Acad. Sci. USA*, **101**, 12916–12921.
28. Buchete, N.-V., Tycko, R. & Hummer, G. (2005). Molecular dynamics simulations of Alzheimer's beta-amyloid protofilaments. *J. Mol. Biol.* **353**, 804–821.
29. Lopez de la Paz, M., de Mori, G. M. S., Serrano, L. & Colombo, G. (2005). Sequence dependence of amyloid fibril formation: insights from molecular dynamics simulations. *J. Mol. Biol.* **349**, 583–596.
30. Hortschansky, P., Schroeckh, V., Christopeit, T., Zandomeneghi, G. & Fändrich, M. (2005). The aggregation kinetics of Alzheimer's beta-amyloid peptide is controlled by stochastic nucleation. *Protein Sci.* **14**, 1753–1759.
31. Christopeit, T., Hortschansky, P., Schroeckh, V., Guhrs, K., Zandomeneghi, G. & Fändrich, M. (2005). Mutagenic analysis of the nucleation propensity of oxidized Alzheimer's beta-amyloid peptide. *Protein Sci.* **14**, 2125–2131.
32. Sabate, R. & Estelrich, J. (2005). Evidence of the existence of micelles in the fibrillogenesis of β-amyloid peptide. *J. Phys. Chem. ser. B*, **109**, 11027–11032.
33. Yong, W., Lomakin, A., Kirkitadze, M. D., Teplow, D. B., Chen, S.-H. & Benedek, G. B. (2002). Structure determination of micelle-like intermediates in amyloid beta -protein fibril assembly by using small angle neutron scattering. *Proc. Natl Acad. Sci. USA*, **99**, 150–154.
34. Prusiner, S. B. (1991). Molecular biology of prion diseases. *Science*, **252**, 1515–1522.
35. Hurshman, A. R., White, J. T., Powers, E. T. & Kelly, J. W. (2004). Transthyretin aggregation under partially denaturing conditions is a downhill polymerization. *Biochemistry*, **43**, 7365–7381.
36. Lomakin, A., Chung, D. S., Benedek, G. B., Kirschner, D. A. & Teplow, D. B. (1996). On the nucleation and growth of amyloid beta-protein fibrils: detection of nuclei and quantitation of rate constants. *Proc. Natl Acad. Sci. USA*, **93**, 1125–1129.
37. Rhoades, E. & Gafni, A. (2003). Micelle formation by a fragment of human islet amyloid polypeptide. *Biophys. J.* **84**, 3480–3487.
38. Serio, T. R., Cashikar, A. G., Kowal, A. S., Sawicki, G. J., Moslehi, J. J., Serpell, L. *et al.* (2000). Nucleated conformational conversion and the replication of conformational information by a prion determinant. *Science*, **289**, 1317–1321.
39. Xu, S., Bevis, B. & Arnsdorf, M. F. (2001). The assembly of amyloidogenic yeast sup35 as assessed by scanning (atomic) force microscopy: an analogy to linear colloidal aggregation? *Biophys. J.* **81**, 446–454.
40. Balbirnie, M., Grothe, R. & Eisenberg, D. (2001). An amyloid-forming peptide from the yeast prion Sup35 reveals a dehydrated β-sheet structure for amyloid. *Proc. Natl Acad. Sci. USA*, **98**, 2375–2380.
41. Nelson, R., Sawaya, M., Balbirnie, M., Grothe, R. & Eisenberg, D. (2005). Structure of the cross-β-spine of amyloid-like fibrils. *Nature*, **435**, 773–778.
42. Reches, M. & Gazit, E. (2005). Self-assembly of peptide nanotubes and amyloid-like structures by charged-termini-capped diphenylalanine peptide analogues. *Israel J. Chem.* **45**, 363–371.
43. Cohen, F. & Kelly, J. W. (2003). Therapeutic approaches to protein-misfolding diseases. *Nature*, **426**, 905–909.
44. Thirumalai, D., Klimov, D. K. & Dima, R. I. (2003). Emerging ideas on the molecular basis of protein and peptide aggregation. *Curr. Opin. Struct. Biol.* **13**, 146–159.
45. Fändrich, M., Forge, V., Buder, K., Kittler, M., Dobson, C. M. & Diekmann, S. (2003). Myoglobin forms amyloid fibrils by association of unfolded polypeptide segments. *Proc. Natl Acad. Sci. USA*, **100**, 15463–15468.
46. Chiti, F., Stefani, M., Taddei, N., Ramponi, G. & Dobson, C. M. (2003). Rationalization of the effects of mutations on peptide and protein aggregation rates. *Nature*, **424**, 805–808.
47. Tartaglia, G. G., Cavalli, A., Pellarin, R. & Caflisch, A. (2004). The role of aromaticity, exposed surface, and

dipole moment in determining protein aggregation rates. *Protein Sci.* **13**, 1939–1941.
48. Hamada, D., Yanagihara, I. & Tsumoto, K. (2004). Engineering amyloidogenicity towards the development of nanofibrillar materials. *Trends Biotechnol.* **22**, 93–97.
49. Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. & Karplus, M. (1983). CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **4**, 187–217.
50. Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, **32**, 241–254.

Edited by F. E. Cohen

(Received 22 February 2006; accepted 6 May 2006)
Available online 5 June 2006

6.2 Interpreting the aggregation kinetics of amyloid peptides. [Supplementary Material of J. Mol. Biol. 2006, *360*, 882].

Supplementary Information

Riccardo Pellarin and Amedeo Caflisch*

Department of Biochemistry
University of Zürich
Winterthurerstrasse 190
CH-8057 Zürich, Switzerland
Phone: (+41 44) 635 55 21
FAX: (+41 44) 635 68 62
email: caflisch@bioc.unizh.ch

* Corresponding author

February 22, 2006

February 22, 2006

Contents

1 Nomenclature

Energy terms

E^{vdW}	Van der Waals energy minimum for hydrophobic spheres
q	Dipole partial charge
E_b	Barrier height
E_π	Potential energy at the protected state π
E_β	Potential energy at the amyloid state β
dE	Potential energy difference between the π and the β states
$P(E_b, dE)$	CMAP dihedral potential
$C_{\phi_0}(\phi)$	Constraining dihedral potential
β-stable	Model with $dE \gtrapprox 0$ dihedral potential
β-unstable	Model with $dE \lessapprox -2.0$ dihedral potential

Monomer states and fibril morphology

π	Protected state
β	Amyloid state
m	Monomeric state
M	Micellar state
f	Monofilament state
F	Fibril state
$F(X_y)$	Fibril morphology symbol

System observables

n_p	Parallel polar contacts
n_h	Hydrophobic contacts

Kinetic observables

t_{lp}	Lag phase time obtained by exponential fitting
t_{50}	Lag phase measured at 50% amplitude (delay time)
k_e	Elongation rate
t_{50}^M	Time of micellization

Concentration and aggregation numbers

C	Total monomer concentration
C_M	Micelle concentration
C_M^r	Critical concentration of micelle formation
C_F^r	Critical concentration of fibril formation
N_T	Total number of simulated monomers
N_m	Number of dissociated monomers
N_M	Micelle aggregation number
N	Oligomer size
N^*	Nucleus size

Thermodynamics

$\Delta G_{\beta\pi}(N)$	Free energy difference between π and β states in an oligomer of size N
$\Delta G_{\beta\pi}(1)$, $\Delta G_{\beta\pi}$	Free energy difference between π and β in the monomeric form

2 Supplementary description of the model

The coarse-grained model developed for studying aggregation kinetics and thermodynamics is a compromise between the mesoscopic detail and the computational efficiency, which are conflicting requirements. Each monomer has internal flexibility and can interact through electrostatics and van der Waals forces with other monomers. Some of the model parameters are chosen to enforce a certain geometry (such as the bonds, the spheres radii and the angles). The strength of van der Waals and electrostatic energy terms are first varied to evaluate the effects on the formation of aggregates. Then the dihedral potential is changed to analyze different aggregation pathways and kinetics. The simplified model does not represent a particular protein; it is useful to understand different aggregation scenarios as observed for different amyloidogenic sequences and experimental conditions.

2.1 The force field

The monomer consists of 10 spherical beads, four of which represent the "backbone" (A2 A3 A6 A10) and six the "sidechains" (A1 A4 A5 A7 A8 A9) (Figure 1). The backbone consists of two identical dipoles with a partial charge q expressed in electronic units; this part of the monomer is designed to interact specifically by intermolecular dipole-dipole interactions. The larger beads represent the sidechains and interact by van der Waals forces. The energy E of the system is evaluated using the following force field formula:

$$\begin{aligned} E \;=\; & \sum_{\text{bonds}} k_b(l-l_0)^2 + \\ & \sum_{\text{angles}} k_a(\theta-\theta_0)^2 + \\ & \sum_{\text{dihedrals}} F(\phi) + \\ & \sum_{i,j} E_{ij}^{vdW}\left[\left(\frac{r_{ij}^{vdW}}{r_{ij}}\right)^{12} - 2\left(\frac{r_{ij}^{vdW}}{r_{ij}}\right)^{6}\right] + \\ & \sum_{i,j} \frac{q_i q_j}{4\pi\varepsilon_0\varepsilon_m r_{ij}} \end{aligned} \tag{1}$$

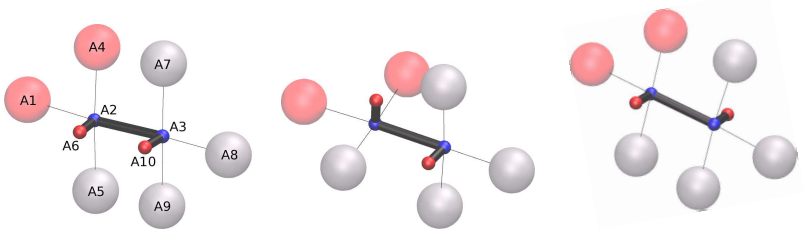


Figure 1: Model of the monomer: light gray spheres are hydrophilic and light red spheres are hydrophobic. The bold black bonds indicate the polar system. Here the positive charges are blue and negative are red. The monomer in the cis state is depicted in the left image, the +90 state in the center, the trans conformation on the right. The spheres drawn here do not reflect the actual van der Waals radii. Geometrical properties and force field parameters are described in the text. Labels indicate the sphere name.

where the sums are evaluated for all bonds, angles, dihedrals and sphere pairs i, j of the system. The variables l, θ and ϕ are the length of the bond, the angle and dihedral values, respectively, while r_{ij} is the distance between the sphere pair i, j. The values of the force constants k_b, k_a and the optimal distances l_0 and angles θ_0 are reported in Table 1. The "molecule" can change its conformation by rotation around the internal dihedral ϕ defined by the beads A6-A2-A3-A10. Depending on the simulation purpose the dihedral potential function $F(\phi)$ is either an harmonic function

$$F(\phi) \equiv C_{\phi_0}(\phi) = k_d(\phi - \phi_0)^2 \tag{2}$$

that restrains the value of the dihedral ϕ around the value ϕ_0, or a potential

$$F(\phi) \equiv P(\phi) \tag{3}$$

designed using the CMAP facility [1], with a grid size of 15 degrees. Several potentials P were investigated in the present work (see also Section 2.3).

The optimal van der Waals energy E^{vdW} and distance r^{vdW} as well as the partial charges q_i are listed in Table 2. The pair constants E_{ij}^{vdW} for the van der Waals interaction in Equation 1 are evaluated using the Lorentz-Berthelot mixing rules [2]. A dielectric constant $\epsilon_m = 1$ is used because the effects of the solvent are taken

Bond energy		
Bead type	k_b $(kcal \cdot mol^{-1} \cdot \mathring{A}^{-2})$	l_0 (Å)
A or C - B	1000.0	5.0
D - B	1000.0	2.0
Angle energy		
Bead type	k_a $(kcal \cdot mol^{-1} \cdot rad^{-2})$	θ_0 (degrees)
A or C or D - B - A or C or D	100.0	90.0

Table 1: Bonding parameter of the force field.

Name	Bead type	E^{vdW} [kcal/mol]	r^{vdW} [Å]	mass [a.u.]	charge q [e.u.]
A1	A	-0.1/-1.6 (∗)	2.5	500	0.0
A2	B	-0.1	2.0	500	0.29/0.52 (∗)
A3	B	-0.1	2.0	500	0.29/0.52 (∗)
A4	A	-0.1/-1.6 (∗)	2.5	500	0.0
A5	C	-0.1	2.5	500	0.0
A6	D	-0.1	2.0	500	-0.29/-0.52 (∗)
A7	C	-0.1	2.5	500	0.0
A8	C	-0.1	2.5	500	0.0
A9	C	-0.1	2.5	500	0.0
A10	D	-0.1	2.0	500	-0.29/-0.52 (∗)

Table 2: Nonbonding parameter of the force field. (∗) Variation of these parameters is investigated in Section 2.4.

into account implicitly by the nonbonding parameters E^{vdW} and q. Assuming an ellipsoidal symmetry, the volume of the monomer is 941 $\mathring{A}^3$, which roughly corresponds to the volume occupied by a peptide of 5 to 11 residues. The mass per bead is set to 500 a.u. This value corresponds to a mass of 4-5 residues, and is chosen to provide stability to the molecular dynamics simulations.

Two types of sidechains are defined: A2, A3 and A5 to A10 have a van der Waals energy minimum E^{vdW} of -0.1 kcal/mol, while A1 and A4, the hydrophobic

sidechains, have a much more favorable van der Waals energy minimum. These two sidechain types generate an "amphipathic" moment which allows the formation of amorphous aggregates such as micelles and the assembly of fibrils. Soreghan *et al.* have emphasized the surfactant properties of β-amyloid peptide and its capability to form solvent oriented structures [3]. For many amyloid proteins, amorphous on and off pathways intermediates have been detected [4–8]. More complex combinations of "sidechain" types could be envisaged for future investigations.

Given the geometry of the monomer (see Section 2.3) and the simplified force field, only two nonbonding parameters are relevant. The van der Waals energy minimum E^{vdW} of the hydrophobic beads A1 and A4 tunes the strength of the non-specific interaction while the partial charge q regulates the strength of the dipole-dipole interaction (see Section 2.4).

2.2 Simulation protocol

Simulations were performed at different temperature values (300-360 K), and concentrations (1.52 - 106.0 mM) using 125 monomers in a box with periodic boundary conditions. The size of the cubic simulation box defines the value of the concentration. A few runs were performed with 1000 monomers to investigate the "seeded" aggregation (see Section 4.2). The simulation protocol is the same for all runs. Monomers are initially placed in a cubic lattice. The system is then heated for 2 ps to the nominal temperature and equilibrated for 20 ps. In this first stage the integration time step is 2 fs and there is no shake constraint. The second stage is a more intensive equilibration; the monomer centers of mass are constrained at their position and simulated for 50 ns. The purpose is to equilibrate the dihedral degree of freedom. The third stage is the production, with previous constraints released. For the second and the third stages the time step is 50 fs, all bonds are restrained with SHAKE [9], and the leapfrog integrator is used for Langevin dynamics at a very low viscosity (0.01 ps^{-1}) which does not influence the thermodynamic properties. For all stages the cutoffs are set to 25 Å for the nonbonding list, 20.0 Å

and 18.0 Å for the nonbonding interactions cutoff and cuton, respectively, with a switching function [10].

2.3 Probing different monomer conformations

The two dipoles are orthogonal to each other in the states π_{90} and π_{-90} corresponding to $\phi = +90$ and -90, respectively. In these conformations the monomers cannot stack along a longitudinal axis, i.e., fibrils cannot be formed. These conformations represent the amyloid-protected state. The states β_0 and β_{180} correspond to cis ($\phi = 0$) and trans ($\phi = 180$) conformation, respectively. These two states can propagate a longitudinal stacking, namely they can form fibrils. The effects of different dihedral conformations were investigated by $1.5\mu s$ runs (125 monomers) with the harmonic potential C defined by the equation 2 (see Figure 2). Besides the π_{90}, π_{-90}, β_0 and β_{180} states mentioned above, conformations with a deviation of ± 15 and ± 30 degrees from cis or trans (noted as $\beta_{\pm 30}$, $\beta_{\pm 15}$, $\beta_{\pm 150}$,$\beta_{\pm 165}$ see Figure 2) were simulated to investigate the effect of monomer chirality on fibril structure. All simulations were performed at aggregation-promoting conditions (see Section 3.4), i.e., concentration of 20.88 mM, 310 K, and aggregation-promoting nonbonding parameters (see Section 2.4), i.e., $E^{vdW} = -1.3$ kcal/mol for the hydrophobic spheres and $q = 0.34$ electronic units.

For all monomeric conformations the system readily starts to form aggregates of different kinds. As mentioned above, conformations π_{90} and π_{-90} do not produce ordered aggregates but rather spherical micellar assemblies (see figure 4) where the hydrophobic spheres are partitioned into the core and the hydrophilic spheres are exposed. These micelles are in equilibrium with dissociated monomers and have an average size of 20-23 monomers (see Section 3.4). All other conformers associate into ordered structures, but interestingly only the chiral conformers produce fibril-like aggregates. The simulations where the monomers are constrained to be either in the state β_0 or β_{180} yield an ordered oligomeric assembly that resembles a disordered crystal lacking a precise cylindrical symmetry (see Figure 2.2) rather

than a fibril. The remaining simulations ($\beta_{\pm 15}, \beta_{\pm 30}, \beta_{\pm 165}, \beta_{\pm 150}$) yield a single cylindrical aggregate consisting of three to four filaments intertwined together, and assembled around the hydrophobic core (see Figure 2.1-3). These ordered aggregates display a twist clockwise or counter-clockwise depending on the chirality (in Figure 2 the circular arrows indicate the helicity of the fibril). These observations agree with a statistical mechanical model for the assembly of chiral molecules [11]. Taken together, the simulation results reveal a complex conformational scenario; in spite of a minimal set of system degrees of freedom, by far smaller than a real polypeptide, the morphological heterogeneity is noticeable. It is important to note that real fibrils do not display a large morphology assortment [12, 13]. In the case of $A\beta_{42}$ only two main morphologies are observed, and probably only two monomer conformations are favored among all possible to form fibrils [13].

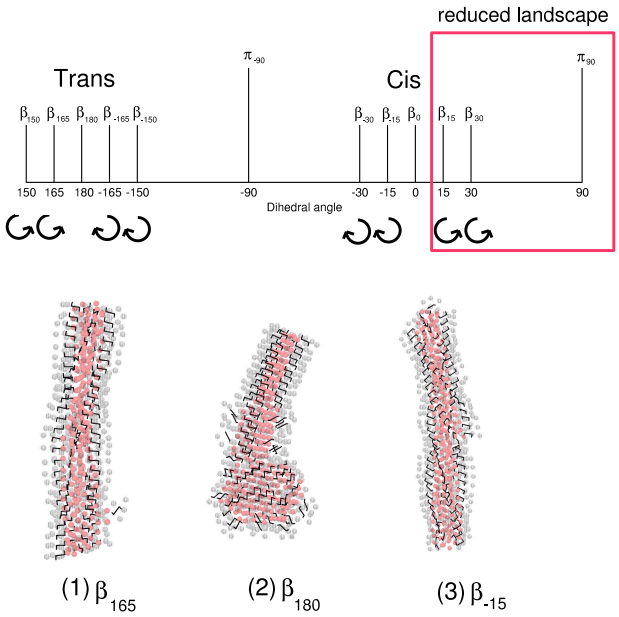


Figure 2: Effects of monomer dihedral angle on aggregation behaviour. Circular arrows indicate the helicity of the resulting fibril. The red box marks the range of dihedral angle values focussed upon by using the CMAP potential (see Figure 3). Fibrils resulting from the simulations of monomers in the state β_{165} (1), β_{180} (2), and β_{-15} (3). Black lines connect the spheres belonging to the polar system, light gray points are the hydrophilic spheres, and light red points are the hydrophobic spheres.

To prevent sampling of redundant, i.e., symmetry-related, conformations a value of 5.0 kcal/mol for all ϕ values outside the interval $7.5 - 97.5$ degrees was imposed by CMAP [1]. This procedure renders all conformations outside such interval inaccessible. Therefore the accessible ϕ-value interval includes the state π_{90}, which is ordered-aggregation protected, and the states β_{15} and β_{30} that form twisted fibrils (see Figure 2). It is now convenient to introduce a short notation for the two conformations: the *β-aggregation protected* state $\pi \equiv \pi_{90}$ and the *β-aggregation competent* state $\beta \equiv \beta_{15}$ or β_{30}.

A dihedral potential function can be introduced in the reduced region to explore different kinetic and thermodynamic properties of the monomer and to investigate their influence on fibril formation. The dihedral potential is defined by two parameters (Figure 3): E_π is the energy of the aggregation protected conformation for ϕ values ranging between 67.5 and 97.5 degrees, while E_b is the energy at the barrier, defined for the $52.5 < \phi < 67.5$ interval.

By fixing the reference state at $E_\beta = 0$ the dihedral potential is fully defined by $P(E_b; dE)$ where E_b is the potential value at the barrier, and $dE = E_\pi - E_\beta$ is the potential difference between the states π and β. As an example $P(1.0; -2.5)$ is

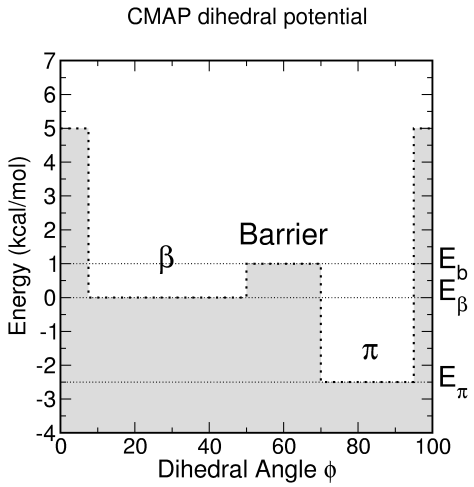


Figure 3: CMAP dihedral potential $P(1.0; -2.5)$. E_β is the energy for the aggregation-competent state, E_b is the energy at the barrier, and E_π is the energy at the protected state.

a potential with a 1.0 kcal/mol barrier for the $\beta \rightarrow \pi$ conversion, and a protected state π stabilized by -2.5 kcal/mol.

2.4 Probing nonbonding parameters

The effects of the variation of the two nonbonding parameters, E^{vdW} and q, are explained in this section. For three different potentials [$P(0;0)$, $P(1.0;-2.5)$ and $C_{+90}(\phi)$] simulations at different values of E^{vdW} for the hydrophobic spheres and q are performed (Table 3). There are different types of ordered and disordered aggregates of increasing complexity: micelles (M), single filament (f), non-twisted bundles of filaments ($F(I)$), twisted bundles of filaments ($F(II)$), and other ordered aggregates ($F(III)$) that cannot be classified in the previous two groups (Figure 4). The bundles (fibrils) can be of different sizes (two to four filaments, noted with a number, see caption of Table 3), but also a single fibril can present segments with a variable number of filaments.

All three investigated potentials show a common feature: the variation of the two parameters defines multiple phase change. At low q and marginally favorable E^{vdW} the monomers are dissociated. This corresponds to the top-right corner of the tables. An equilibrium of monomers with oligomers at the transition points is often observed, or eventually micelles with fibrils (as in the case of potential $P(1.0;-2.5)$, $q = 0.31$ and $E^{vdW} = -1.6$). At the bottom-left corner no coexistence is observed (pure ordered aggregates).

The restrained potential. The simulations performed with the $C_{+90}(\phi)$ potential are control simulations. As mentioned above, the π state ($\phi = +90$) cannot form any ordered aggregate but only micelles. As it is evident from Table 3, micelles formation occurs for $E^{vdW} \leq -1.0$ kcal/mol, weakly depending on the value of the charge q.

The β-unstable potential $P(1.0;-2.5)$. Favorable values of E^{vdW} and high values of q promote the ordered aggregation. $F(I)$ fibrils are observed at high q, together with single filaments f. The size of the bundle increases with more

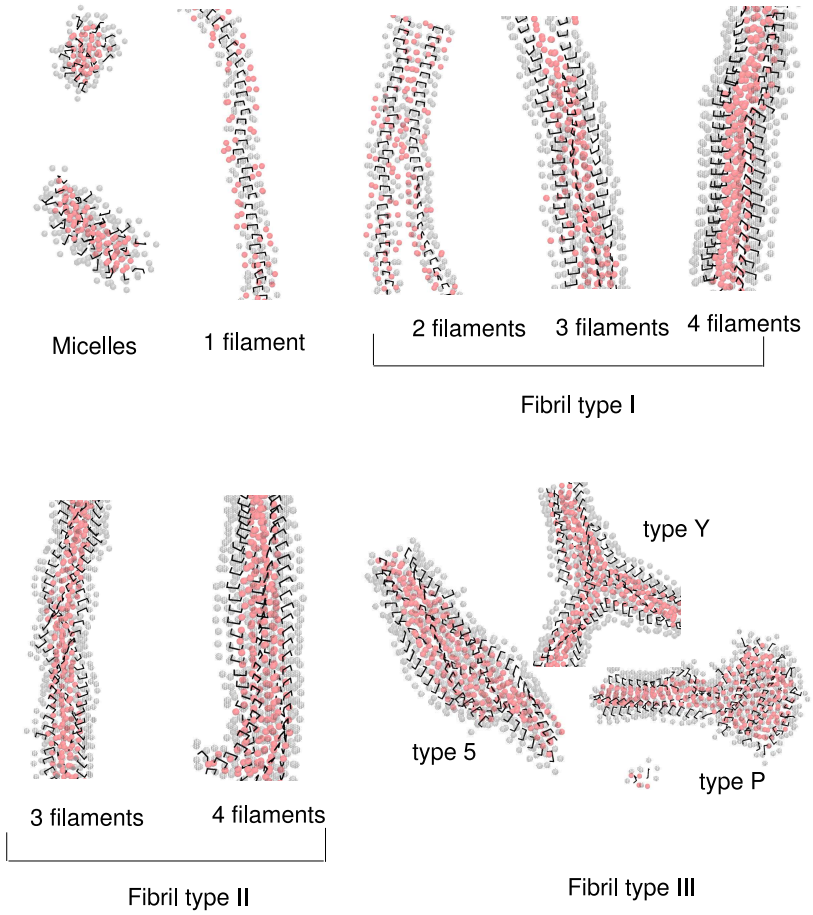


Figure 4: Different aggregation morphologies observed in simulations. The dipole system is indicated in black, the hydrophilic spheres are indicated in transparent gray, and the hydrophobic spheres are light red. Micelles are non-ordered spherical or elliptical metastable aggregates. Fibrils are multi-filament assemblies. Fibrils of "type I" are not twisted, while fibrils of "type II" are twisted. Fibrils of "type III" include everything that cannot be classified in the previous types: type III-5 is a bundle of 5 filaments, III-y is a cross of two or more fibrils, and type III-p is a planar fibril similar to a ribbon.

favorable E^{vdW}, i.e., by increasing the hydrophobicity. Twisted fibrils $F(II)$ are present at lower q, indicating that the twisting is a result of balancing of these two parameters.

The β-stable $P(0;0)$ potential. The region where the fibrillation occurs is larger than in the $P(1.0;-2.5)$ case, which is a consequence of the larger thermo-

Potential $C_{+90}(\phi)$

E^{vdW}	q=0.52	0.45	0.4	0.36	0.34	0.31	0.29
-0.1	m	m	m	m	m	m	m
-0.4	m	m	m	m	m	m	m
-0.7	m	m	m	m	m	m	m
-1.0	m+M	m+M	m	m	m	m	m
-1.3	M	M	m+M	m+M	m+M	m+M	m+M
-1.6	M	M	M	M	M	M	M

Potential $P(1.0;-2.5)$

E^{vdW}	q=0.52	0.45	0.4	0.36	0.34	0.31	0.29
-0.1	m+f:1	m	m	m	m	m	m
-0.4	F(I_2):1	m	m	m	m	m	m
-0.7	F(I_3):1	m	m	m	m	m	m
-1.0	F(I_3):2	F(II_3):2	m	m	m	m	m
-1.3	F(I_4):3	F(II_{3-4}):3	F(III_P):3	F(II_4):3	m+F(II_4):3	m+M	m+M
-1.6	F(III_P):3	F(II_{3-4}):4	F(III_Y):4	F(III_P):4	F(II_{3-4}):4	M+F(II_4):4	M

Potential $P(0;0)$

E^{vdW}	q=0.52	0.45	0.4	0.36	0.34	0.31	0.29
-0.1	f:1	m+f:1	m	m	m	m	m
-0.4	F(I_{2-3}):1	m+F(I_3):1	m+f:1	m	m	m	m
-0.7	F(I_{4-5}):1	m+F(I_3):1	m+F(I_3):1	m	m	m	m
-1.0	F(I_{2-4}):1	F(I_3):1	m+F(II_{3-4}):1	m+F(II_4):2	m+F(II_4):2	m+F(II_4):2	m+F(II_4):2
-1.3	F(II_{3-4}):1	F(I_{2-4}):1	F(II_4):2	F(III_5):2	m+F(II_4):3'	m+F(II_3):3'	m+F(II_{3-4}):3'
-1.6	F(III_P):2	F(III_P):2	F(III_P):2	F(II_{3-4}):3'	F(II_{3-4}):3'	F($II_3; III_P$):3'	F(II_{3-4}):3'

Table 3: Effect of variation of the hydrophobic strength (E^{vdW} is the van der Waals potential energy well of spheres A1 and A4) and the charge q. Legend: (m) monomers, (M) micelles, (f) filaments, (F) fibrils. For the fibrils it is notated the type and the number of filaments: e.g. F(II_{3-4}) is a fibril of type II with 3 to 4 filaments. The number after the colon is the type of pathway followed by the simulation (see Figure 5). The sign plus "+" indicates coexistence of different phases. The concentration is 20.88 mM and the temperature is 310 K. The colors of the fields stand for the main aggregation morphology: green for micelles, blue for single filaments f, pink for $F(I)$, orange for $F(II)$, and red for $F(III)$.

dynamic accessibility of the β state. It is worth noting that chemically or mutation denaturated proteins are more susceptible to form fibrils than their standard conditions or wild type counterparts, respectively [14–16].

The aggregation process can be monitored by the number of parallel polar contacts and hydrophobic contacts along the trajectories. A parallel polar contact is formed whenever sphere 6 and 2 or sphere 10 and 3 of different monomers are closer than 5 Å. This selection of contacts defines the parallel aggregation, whereas the antiparallel is not observed because of the amphipathicity of the monomer (see above). A hydrophobic contact is formed whenever spheres 1 or 4 of different monomers are closer than 5 Å. The total number of polar n_p and hydrophobic n_h contacts is evaluated for each frame:

$$\begin{aligned} n_p &= \frac{1}{2}\sum_{i=A6}\sum_{j=A2}\delta(r_{ij}\leq 5) + \\ &\quad \frac{1}{2}\sum_{i=A10}\sum_{j=A3}\delta(r_{ij}\leq 5) \qquad (4) \\ n_h &= \frac{1}{2}\sum_{i=A1,A4}\sum_{j=A1,A4}\delta(r_{ij}\leq 5) \qquad (5) \end{aligned}$$

where, for instance, the summation index $i = A1, A4$ runs for all A1 and A4 spheres, and the function δ is equal to 1 if the distance between the two spheres r_{ij} is less than 5 Å. These quantities are used as progress variables of the aggregation process. A point in Figure 5 represents values of n_p and n_h of a single snapshot. Fibril formation corresponds to a trace of points that spans from the origin (monomeric state) to a maximal value of both variables that is around 225 for n_p and 600 for n_h in the simulation with 125 monomers. Given the geometry and size of the monomer the number of parallel polar and hydrophobic contacts per monomer incorporated into a fibril is about 2 and 5, respectively.

The left and right plot of figure 5 display the (n_p, n_h)-values of the snapshots saved along the simulations with the $P(0;0)$ and $P(1.0;-2.5)$ potentials respectively. Different pathways of fibril formation can be identified. The meaning of the

different pathways can be understood considering first the path number 2 which follows a straight line for both potential models. The variables under consideration depend linearly on each other in path 2, namely for any association event n_p, and n_h increase by a value that is constant along the aggregation process. The physical meaning of this behavior is that the fibril, once it has nucleated, progressively increases its size by absorbing monomers or small oligomers. During elongation the morphology of the progressing fibril it is similar to the final fibril.

There are different considerations for path 1, path 3-3' and path 4. Path 1 is a double stage transition for both potentials. In the first stage monomers form aggregates that maximize the number of polar interactions. These oligomers assemble by increasing hydrophobic contacts number in the second stage. They can be identified with single filaments. Only for the $P(1.0; -2.5)$ potential, paths of type 3 and 4 are observed. Clouds of points in the low n_p and high n_h can be interpretated as disordered aggregates. From visual examination, these on-pathway assemblies are micellar-like oligomers very similar to those obtained for the $C_{+90}(\phi)$ potential.

In the case of path 3' of $P(0; 0)$ potential, these on-pathway micellar aggregates are not present: there is a faster transition from monomer state to fibril state. Path 3' is thus different from path 3. The pathway 4 is observed only at high hydrophobic strength; it is similar to pathway 3, but shifted towards the high hydrophobic contacts content. In other words it is a nucleation from a bigger micellar aggregate. Likewise the pathway 3' is present at high hydrophobic strength for the $P(0; 0)$ potential. As mentioned above the micellar state is absent for this pathway, it is consequently equivalent to path 2, shifted towards high content of hydrophobic contacts. Pathways 3 and 4 are therefore qualitatively the same, and the same is valid for pathways 2 and 3'. This allows a characterization of the fibril formation pathways in three main classes (see Figure 6).

The path numbers are reported in Table 3 for all parameters pairs. The preferred path depends on the values of q and E^{vdW}. For potential $P(1.0; -2.5)$ the

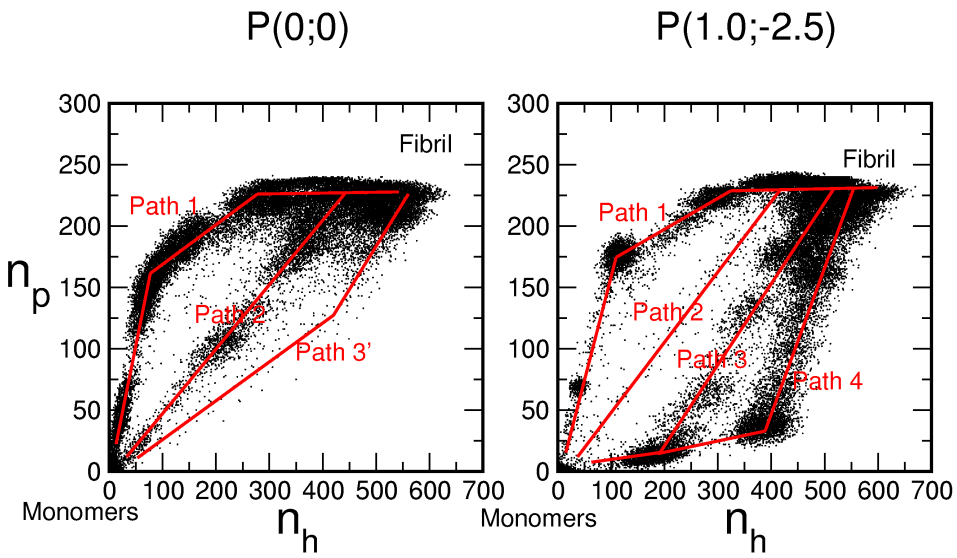


Figure 5: Pathways of fibril formations. The number of parallel polar interactions n_p and the number of hydrophobic interactions n_h for all trajectories defined in Table 3 for $P(0;0)$ and $P(1.0;-2.5)$ potentials. The clustering of points permits a classification of diverse aggregation pathways. Red lines are meant to guide the eyes.

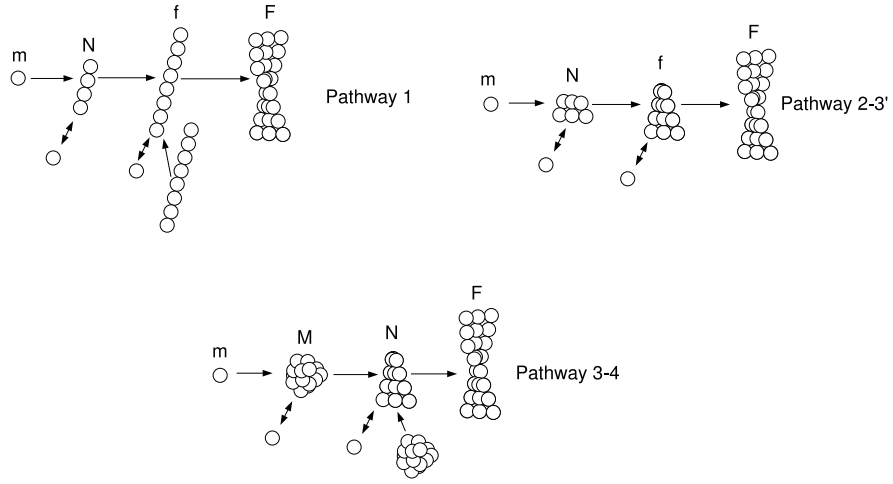


Figure 6: Pathways classification. Legend: (m) monomer, (M) micelle, (N) nucleus, (f) filament, (F) fibril.

predominant path changes from 1 to 4 by increasing the strength of the hydrophobic contacts. For the $P(0;0)$ potential the variation is influenced by both parameters.

These results allow us to describe the distinctive aggregation pathways for the

two potential models. For the β-**unstable** model $P(1.0; -2.5)$ the micellar state is an intermediate required for nucleating the ordered aggregation. Especially when the monomer specific affinity is low, i.e., when the charge q is small, fibril formation occurs via micellar intermediates. The hydrophobic interactions are essential for the fibril nucleation step. For the β-**stable** $P(0; 0)$ potential the structure interconversion is kinetically fast, the monomers depositing onto an ordered aggregate promptly convert its conformation to the β state. The elongation process is driven by polar interactions and hydrophobic interactions play a role only in the fibril morphology, i.e., filaments assembly.

With the values of $E^{vdW} = -1.3$ kcal/mol and $q = 0.34$ e.u. an equilibrium between monomers and fibrils is reached at the final stage of the β-unstable $P(1.0; -2.5)$ potential simulations (Table 3). This thermodynamic behavior is an important feature of a realistic model system. For this reason these values are adopted for all kinetic and thermodynamic analysis in the following sections and the main text.

3 Supplementary methods

In this section, and the main text, values of $E^{vdW} = -1.3$ kcal/mol and $q = 0.34$ e.u. are used for the van der Waals energy minimum and the partial charge, respectively.

3.1 Evaluation of system kinetics

The lag phase time t_{lp} is extracted from the exponential fitting of the time series of the number of parallel polar contacts n_p (defined by Equation 4). The fitting function is

$$n_p(t) = n_p(0) + [n_p(\infty) - n_p(0)](1 - e^{-k_e(t-t_{lp})})S(t, t_{lp}) \quad (6)$$

where $n_p(0)$ is the initial number of polar contacts, $n_p(\infty)$ the equilibrium value, and k_e the elongation rate. $S(t, t_{lp})$ is a switching function that is 0 for $t < t_{lp}$ and 1 for $t \geq t_{lp}$ and is needed to force the function $n_p(t)$ to have a constant value for $t < t_{lp}$ (Figure 7). The **lag phase** lasts during $t < t_{lp}$ while the **final fibril-monomer equilibrium** is established at times $t > 10t_e$ when the function $n_p(t)$ exceeds 90% of its maximal value $n_p(\infty)$, where $t_e = 1/k_e$. The **elongation** occurs during $t_{lp} \leq t \leq 10t_e$.

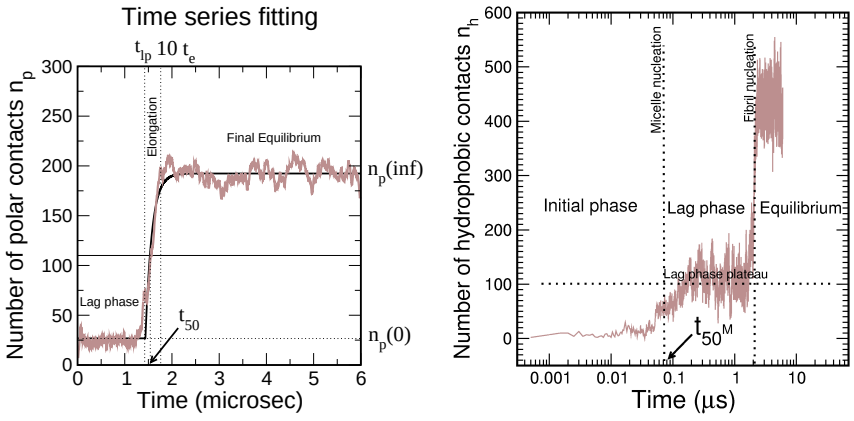


Figure 7: Example of the fitting of a time series of the number of parallel polar contacts n_p (left). Time series of the number of hydrophobic contacts n_h and the time of micelle formation t_{50}^M (right).

For the kinetic investigations of Sections 3.7 and 3.8 it is more appropriate to use a slightly different definition of the lag phase time, i.e., the time needed to reach 50% of the maximal amplitude t_{50} [17] (Figure 7). The t_{50} (termed delay time henceforth) is more robust than the lag phase time t_{lp}, especially for nucleation events with a short lag time. The lag phase time t_{lp} is used only for an exact measure of the lag phase, e.g. when thermodynamic properties at the lag phase are investigated (Sections 3.3, 3.4 and 3.5).

The number of parallel polar contacts n_p is not appropriate to monitor the nucleation kinetics of disordered aggregates such as micelles. For this purpose it is convenient to use the number of hydrophobic contacts n_h (defined in Equation 5). In analogy with t_{50}, one can define the *time of micellization* t_{50}^M as the time needed to reach 50% of lag phase plateau amplitude starting from $t = 0$ (see Figure 7). The t_{50}^M time can be evaluated only for β-unstable models and at low concentration, where the lag phase is long enough to separate the micellization from the fibril nucleation phase. The average value of t_{50}^M is 30 ns at C=8.5 mM (Figure 2 in the main text).

3.2 Clustering

A clustering algorithm is used to calculate the size of the oligomeric species along the simulations. It is based on the matrix of contacts D_{ij} between labeled monomers. Given a single frame of the simulation D_{ij} is equal to one if any sphere of monomer i is closer than 6.0 Å to any sphere of monomer j, otherwise it is zero. D_{ij} is equivalent to the first neighbor matrix, $d_{ij}^{(1)}$. The second neighbor matrix $d^{(2)}$ is constructed from $d^{(1)}$ including the neighbors of the first neighbors. The converged contact matrix $d_{ij}^{(\infty)}$ is defined by the following recursive sequence

$$
\begin{aligned}
d_{ij}^{(1)} &= D_{ij} \\
d_{ij}^{(n)} &= \begin{cases} 1 \quad \text{if} \quad d_{ij}^{(n-1)} = 1 \\ 1 \quad \text{if} \quad d_{ik}^{(n-1)} = 1 \qquad \text{and} \qquad d_{kj}^{(n-1)} = 1 \\ 0 \quad \text{otherwise} \end{cases} \\
d_{ij}^{(\infty)} &= \lim_{n \to \infty} d_{ij}^{(n)}
\end{aligned}
\tag{7}
$$

which yields a block matrix of ones and zeroes. Each block represents a cluster of monomers and contains first-neighbors, second neighbors, third and so on. This procedure is equivalent to a hierarchical clustering performed with a spanning tree technique [18]. With the converged contact matrix one can identify clusters (i.e. tagging each oligomer with an identification number), list monomers belonging to a specific oligomer and make statistics on the size of oligomers.

3.3 Cluster size histogram

The above definition of oligomeric species allows the statistical analysis of cluster size. The probability that a monomer is aggregated in a cluster of size N is:

$$
p(N) = \langle \frac{1}{N_T} \sum_{i=1,N_T} \delta_{i,t}(N) \rangle_t \tag{8}
$$

where N_T is the total number of simulated monomers, $\delta_{i,t}(N)$ is equal to 1 if the monomer i at time t is embedded in a cluster of size N, and the angular brackets are the time average. This function (termed cluster size distribution) can be evaluated for the lag phase or the final monomer-fibril equilibrium. The elongation phase cannot be analyzed by $p(N)$ being an out of equilibrium dynamic process.

The peaks of the $p(N)$ distribution can be interpreted as stable oligomeric species. The monomer peak ranges from $N = 1$ to 7, the micellar peak from $N = 8$ to 60, and the fibril peak from $N = 61$ to 125 (Figure 8). The height of the peaks depends on the relative stability of the β-competent state as well as the total monomer concentration (Figure 8). For the β-stable potential $P(0;0)$ the

micelle peak is not observed at any concentration value. With incresing concentration the monomer and micelle peak distributions are skewed towards high N values because multi-monomer collisions and multi-micellar collisions, respectively, transiently generate oligomers of a larger size.

3.4 Phase diagrams and critical concentrations

Phase diagrams of temperature T and concentration C were calculated only for the potential $P(1.0; -2.5)$ because of their computational demand (about 2 weeks on 80 CPUs). The probability of a monomer being in the monomeric (m), micellar (M) and fibrillar (F) states are evaluated as cumulative sum of the probability $p(N)$:

$$p_m = \sum_{N=1}^{7} p(N) \tag{9}$$

$$p_M = \sum_{N=8}^{60} p(N) \tag{10}$$

$$p_F = \sum_{N=61}^{125} p(N) \tag{11}$$

In the simulated system three possible phases can coexist: monomeric, micellar (disordered oligomer), and fibril (Figure 9). The results are robust for a monomer-micelle threshold in the range 5-10 and a micelle-fibril threshold between 50 and 70. The C, T-diagram of the lag phase is similar to the one of the final equilibrium for large T values, where the fibrillization is inhibited. At equilibrium one has a triphasic diagram.

The probability distribution $p(N)$ can be used to evaluate the critical concentration of micelle formation C_M^r. The micelle aggregation number in the lag phase is defined as

$$N_M = \sum_{N=8}^{60} N p^{lp}(N) \tag{12}$$

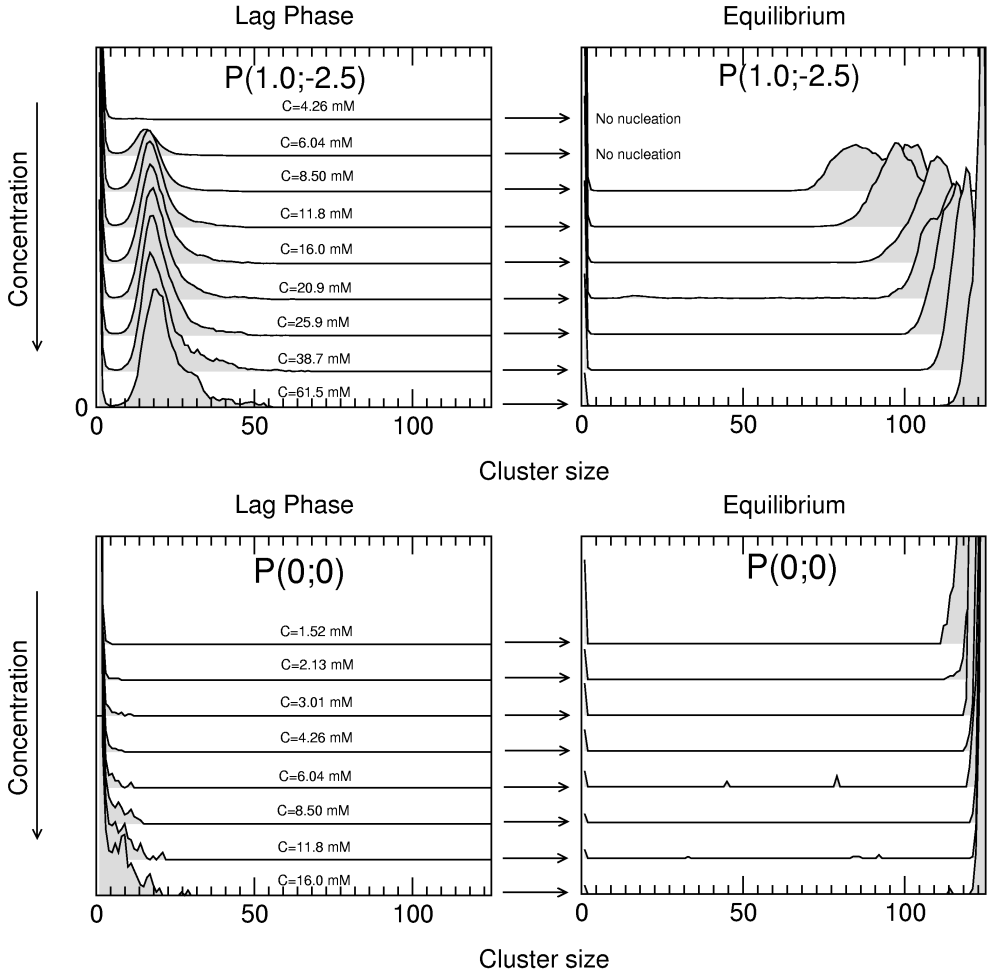


Figure 8: Cluster size histograms of the $P(1.0; -2.5)$ potential (top) and $P(0; 0)$ potential (bottom) calculated in the lag phase (left) and the final equilibrium (right). Histograms belonging to the same concentration are reported in the same row. The z-dimension represent the relative probability.

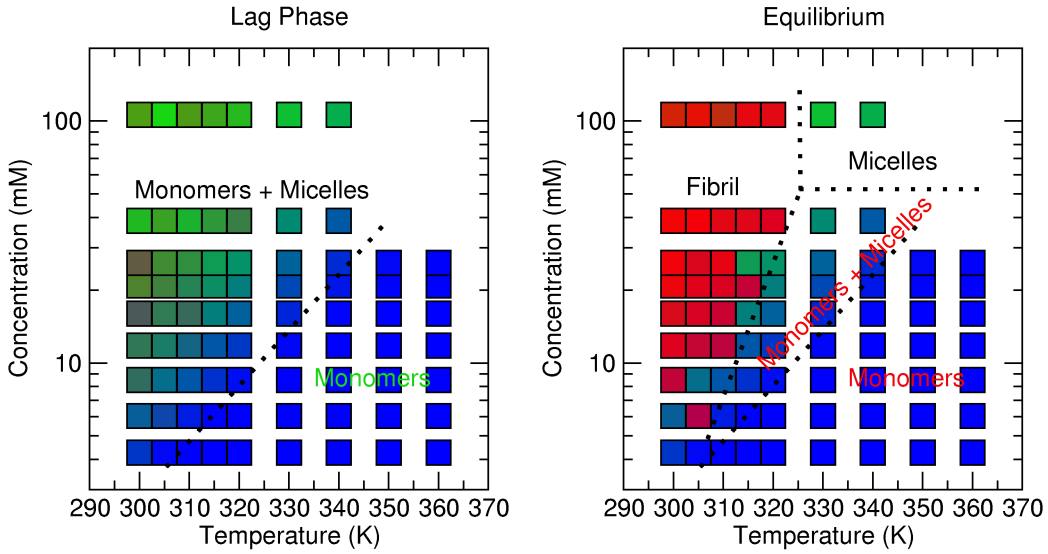


Figure 9: Phase diagram of the lag phase (left) and final equilibrium (right) for the potential $P(1.0; -2.5)$. Each color of the phase diagram is obtained by mixing the red, blue and green components according to the values of the calculated probabilities, i.e., red=p_F, blue=p_m, green=p_M (see Equations 9, 10 and 11).

where p^{lp} is the probability function evaluated only in the lag phase (where there is coexistence of micelles and monomers without fibrils). The number of micelles per simulation box is $N_T p_M N_M^{-1}$, where the total number of simulated monomers N_T is 125. The micelle concentration C_M is derived from the number of micelles in the simulation volume. The micelle aggregation number N_M and concentration are plotted in Figure 10 as a function of the total monomer concentration C for the potential $P(1; -2.5)$ in the lag phase. By extrapolating a linear fit of the concentration of micelles the critical concentration of micelle formation C_M^r can be evaluated. The value is C_M^r =4.36 mM.

Another important observable is the critical concentration of fibril formation C_F^r, that is obtained from the concentration of dispersed monomers in equilibrium with the final fibril. The number of dispersed monomers in the simulation box at the equilibrium phase is:

$$N_m = N_T p_m^{eq} = N_T \sum_{N=1}^{7} p^{eq}(N) \tag{13}$$

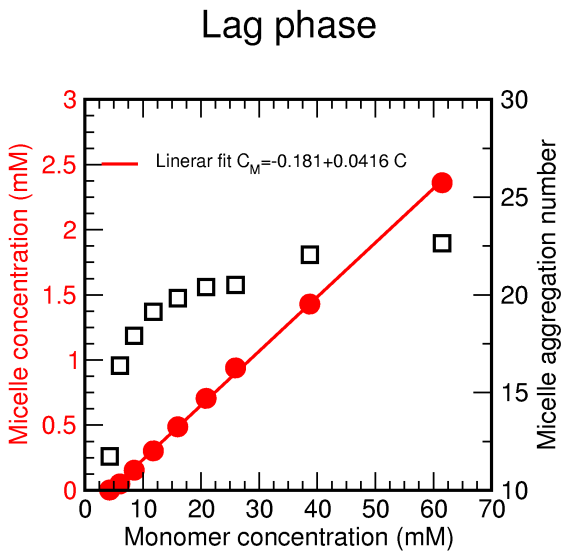


Figure 10: Lag phase of the P(1.0;-2.5) potential. Micelle concentration C_M (red circles) and the micelle aggregation number N_M (squares) as function of the total concentration of monomers C. The straight line is a linear fit whose parameters are reported in the graph.

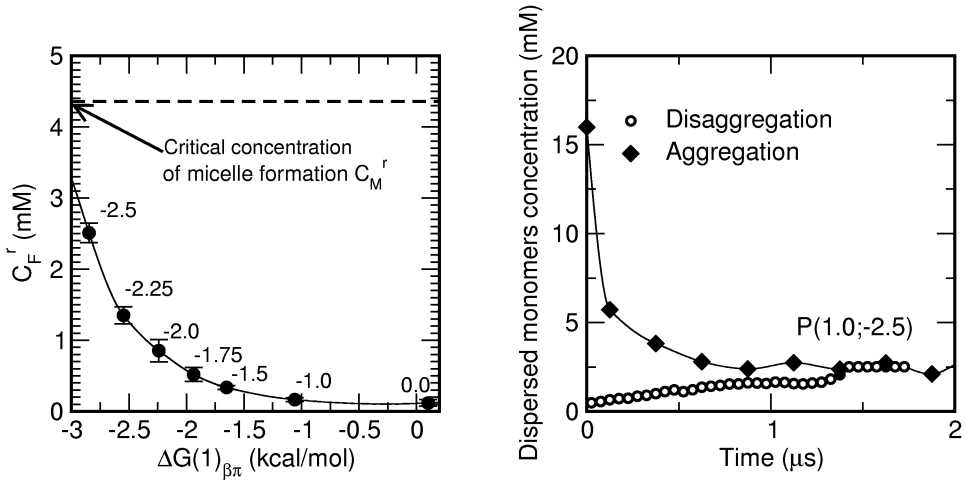


Figure 11: (Left) critical concentration of fibril formation C_F^r as a function of the monomeric $\Delta G_{\beta\pi}(1)$. The numbers displayed near the data points are the value of dE for the potential $P(1.0; dE)$. The critical concentration of micelle formation is indicated by a dashed line. (Right) Validation of the critical concentration of fibril formation, for the potential $P(1.0; -2.5)$. The dispersed monomer concentration is monitored along time for the aggregation process (filled diamonds) and along a simulation of disaggregation started from a previously formed fibril (empty circles).

where p^{eq} is the probability function evaluated at the equilibrium (where there is the coexistence of monomers and fibrils). The value of C_F^r can be evaluated by dividing N_m by the simulation volume. In Figure 11, C_F^r is displayed for different monomer potentials P and plotted against relative stability of the protected state $\Delta G_{\pi\beta}(1)$ (Table 4). Smaller differences in free energy result in lower critical concentration. In other words β-stable models are more reactive and shift the reaction towards the fibril formation. The critical concentration of micelle formation is always higher than the critical concentration of fibril formation; for this reason the micelles disappear at the monomer-fibril equilibrium.

The critical concentration of fibril formation is validated by additional simulations (Figure 11 right plot). The dispersed monomer concentration is evaluated dynamically for an aggregation trajectory (potential $P(1.0; -2.5)$, C=16.0 mM). The final concentration is equal to the predicted critical concentration of 2.5 mM (Figure 11 left plot, dE=-2.5). The reverse reaction is also performed to test if such C_F^r value is approachable also from the disaggregation direction. A fibril, previously prepared at C=16.0 mM, is simulated at the critical concentration of fibril formation (C=C_F^r=2.5 mM). The fibril progressively disassembles and the dispersed monomer concentration increases to the value of 2.5 mM. The same combination of forward and reverse reactions were used to experimentally test the robustness of the critical concentration of fibril formation for the β-amyloid peptide [19].

3.5 Aggregation process

Six monomeric states can be defined as a combination of β or π, and the three oligomeric species m, M, or F: $\pi m, \beta m, \pi M, \beta M, \pi F, \beta F$. As an example πF is the state for a monomer in the π conformation within a fibril. The transition matrix $T_{ij}(\Delta t)$ is defined as

$$T_{ij}(\Delta t) = p(i|j, \Delta t) \tag{14}$$

where i and j are two of the six states and $p(i|j, \Delta t)$ is the conditioned probability

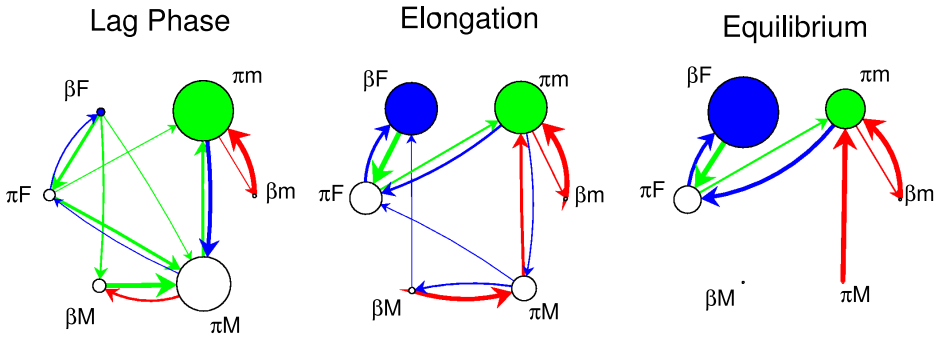


Figure 12: Simplified network representation of the transition matrix T_{ij} in the lag phase, elongation and equilibrium time regimes. The size of the nodes represents the state self transition T_{ii}, and the size of the links is the cross transitions T_{ij} with $i \neq j$. Assuming the initial state πm (green) and the final state βF (blue), blue arrows indicate the pathways leading form πm to βF and green arrows from βF to πm. Only transitions with probability greater than 0.05 are showed. The transition matrix is evaluated on the potential $P(1; -2.5)$ at concentration 11.8 mM and on 15 independent simulations of 12μs each.

of jumping to the j state from state i in a time Δt. The time Δt is chosen as the smallest available time in the simulation (the time of coordinate saving, 0.5 ns) to resolve the fastest events. From the simplified network representation of the transition matrix (Figure 12) it is clear that for the potential model $P(1.0; -2.5)$ the lag phase consists of a micellar association equilibrium $\pi m \leftrightarrow \pi M$ as well as intramonomer interconversions $\pi m \leftrightarrow \beta m$ and $\pi M \leftrightarrow \beta M$. Furthermore, in the lag phase the fibril state F is mainly accessible through the micellar state M. Therefore, in this time regime the fibril state is a transient ordered oligomer in equilibrium with micelles. At the final equilibrium micelles are very unstable and the main pathway is $\beta m \rightarrow \pi m \leftrightarrow \pi F \leftrightarrow \beta F$, indicating that monomers are attached to the fibril in the protected conformation π before assuming the amyloid conformation β. Furthermore, the βm state is off-pathway. This mechanism is consistent with kinetic experiments on radiolabeled $A\beta 40$ peptides where the transition from soluble to amyloid-like conformation of the peptide was suggested to be mediated by interaction with the fibril template (dock-lock mechanism) [20]. The equilibrium $\pi F \leftrightarrow \beta F$ reflects coexistence of monomers π and monomers β in the fibril.

From trajectory visualization it is clear that the conformations β and π populate different domains of the fibril; monomers β are found mainly in the central region of the fibril, whereas monomers π populate the disordered caps. Therefore, isolated monomers in equilibrium with the fibril are continously attaching to and detaching from the caps of the fibril in the β-protected conformation π. The growing phase is regulated by monomer addition, rather than oligomer addition. Collins *et al.* [21], using a combination of kinetic measures, reported a monomer addition growing for the yeast prion.

3.6 $\pi - \beta$ free energy difference

The aggregation number N is a natural progress variable to monitor the polymerization progress of an oligomer. The clustering procedure introduced in Section 3.2, can be used to calculate the free energy difference between the state π and the state β of a monomer belonging to an oligomer of size N:

$$\Delta G_{\beta\pi}(N) = G_{\pi}(N) - G_{\beta}(N) = -\langle kT \log \left(\frac{N_{\pi}(N)}{N_{\beta}(N)} \right) \rangle_N \qquad (15)$$

where $N_{\pi}(N)$ and $N_{\beta}(N)$ are the number of π-monomers and β-monomers, respectively, present in an oligomer of size N, and the angular bracket is the average over all oligomers of size N. This function of the number of monomers does not depend on the concentration because it is an intrinsic property of the oligomer and independent of the surrounding environment (Figure 13.A-B). The lack of dependence on concentration allows the evaluation of the function $\Delta G_{\beta\pi}$ using simulation data at different values of concentration (Figure 13.C).

The models with dihedral energy difference $dE < -2.5$ kcal/mol are not observed to nucleate, even at high concentration (C=61 mM) and long simulation times (16 μs); their aggregation number (or cluster size) N does not exceed $60-70$ (empty circles in Figure 13.C). The $dE = -2.5$ and -2.25 kcal/mol potentials, which are the most β-unstable potentials still capable of fibril formation, have a $\Delta G_{\beta\pi}(N) = 0$ at $N \approx 45$ and 30, respectively. Interestingly, these values roughly

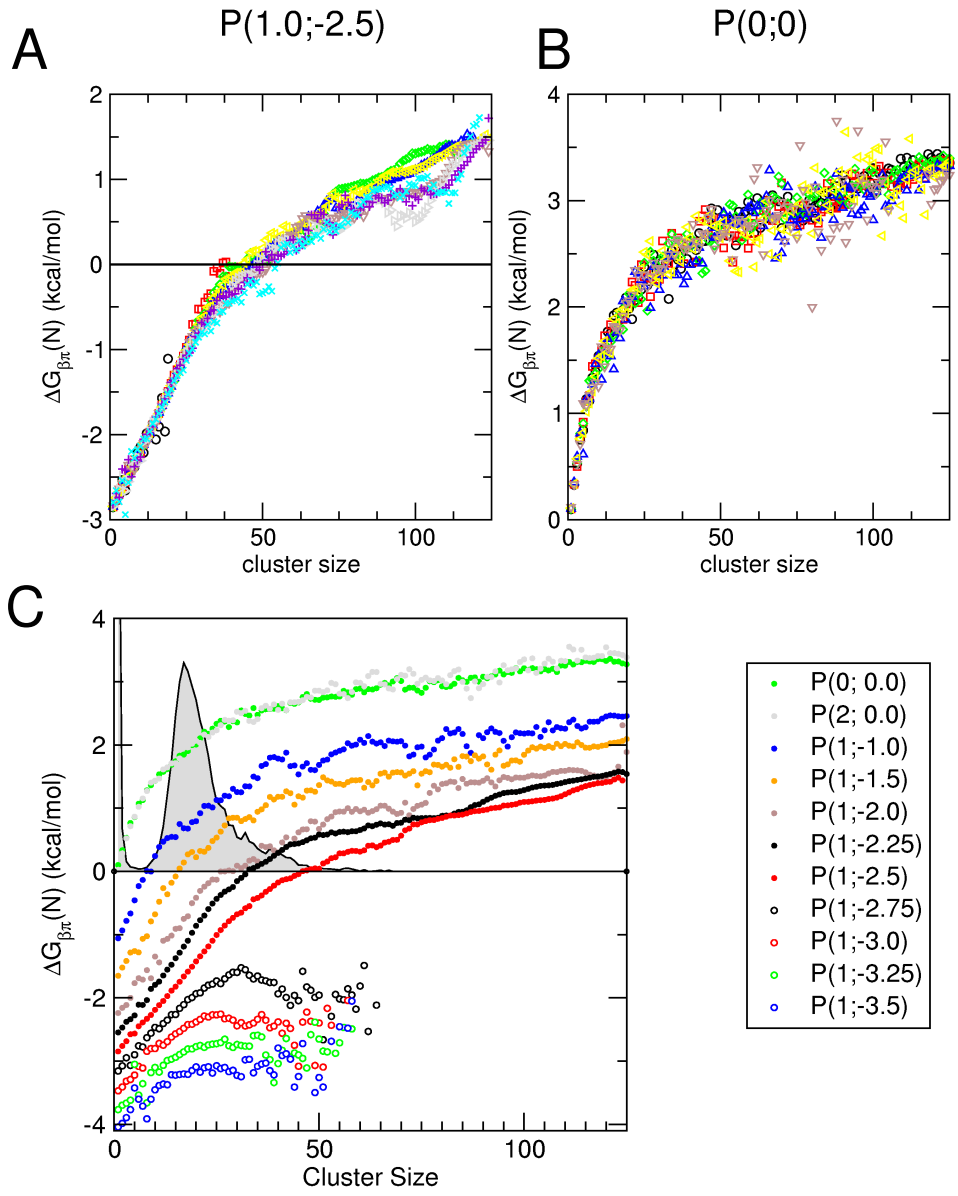


Figure 13: $\Delta G_{\beta\pi}(N)$ evaluated at different concentrations (different colors and symbols) for the potential $P(1;-2.5)$ (A) and $P(0;0)$ (B). (C) Variation of the $\Delta G_{\beta\pi}(N)$ evaluated for different potentials. The gray area in the background is the normalized cluster distribution during the lag phase.

Barrier height	
Potential	$\Delta G_{\beta\pi}(1)$ (kcal/mol)
$P(4;0)$	-0.165
$P(3;0)$	-0.0776
$P(2;0)$	-0.0255
$P(1;0)$	0.0844
$P(0;0)$	0.101

$\pi - \beta$ **Relative stability**	
Potential	$\Delta G_{\beta\pi}(1)$ (kcal/mol)
$P(1;-0.5)$	-0.452
$P(1;-1.0)$	-1.05
$P(1;-1.5)$	-1.65
$P(1;-1.75)$	-1.94
$P(1;-2.0)$	-2.24
$P(1;-2.25)$	-2.54
$P(1;-2.5)$	-2.84
$P(1;-2.75)$	-3.16
$P(1;-3.0)$	-3.47
$P(1;-3.25)$	-3.77
$P(1;-3.5)$	-4.04

Table 4: Free energy difference of isolated monomers $\Delta G_{\beta\pi}(1)$ for all investigated potential models.

correspond to the estimated nucleus sizes of 40 for the $dE = -2.5$ model and 27 for $dE = -2.25$ (see Figure 6 of the main text) indicating that the oligomeric size N_0 at which $\Delta G_{\beta\pi}(N_0) = 0$ identifies thermodynamically the nucleus size, in a way which is consistent with the probabilistic definition of nucleus exposed in Section 3.11. Comparing the $\Delta G_{\beta\pi}(N)$ and the cluster size distribution at lag phase (Figure 13.C), it is revealed that the values of N_0 for nucleating models are in the range where the cluster size distribution has a statistically significant probability. For nucleating β-unstable models $dE = -2.5$ and -2.25 kcal/mol the N_0 values are located at the micelle right tail, indicating that the nucleation step is initiated in an oligomer with a size larger than a micelle. The values of the monomeric free energy difference $\Delta G_{\beta\pi}(1)$ show a shift to more pronounced stabilization of the π state with slightly more negative values than dE (Table 4).

3.7 Concentration influence on kinetics of fibril formation

The time series of the number of parallel polar contacts n_p for different potential models and at different concentration values are shown in Figure 15. The analysis of the concentration dependence of the delay time t_{50} and the elongation rate k_e is reported in Figure 14 C-D. Potentials $P(1;-2.5)$, $P(1;-2.25)$, $P(2;0)$ and $P(0;0)$ were analyzed. Elongation rate k_e is evaluated by fitting with the Equation 6 the

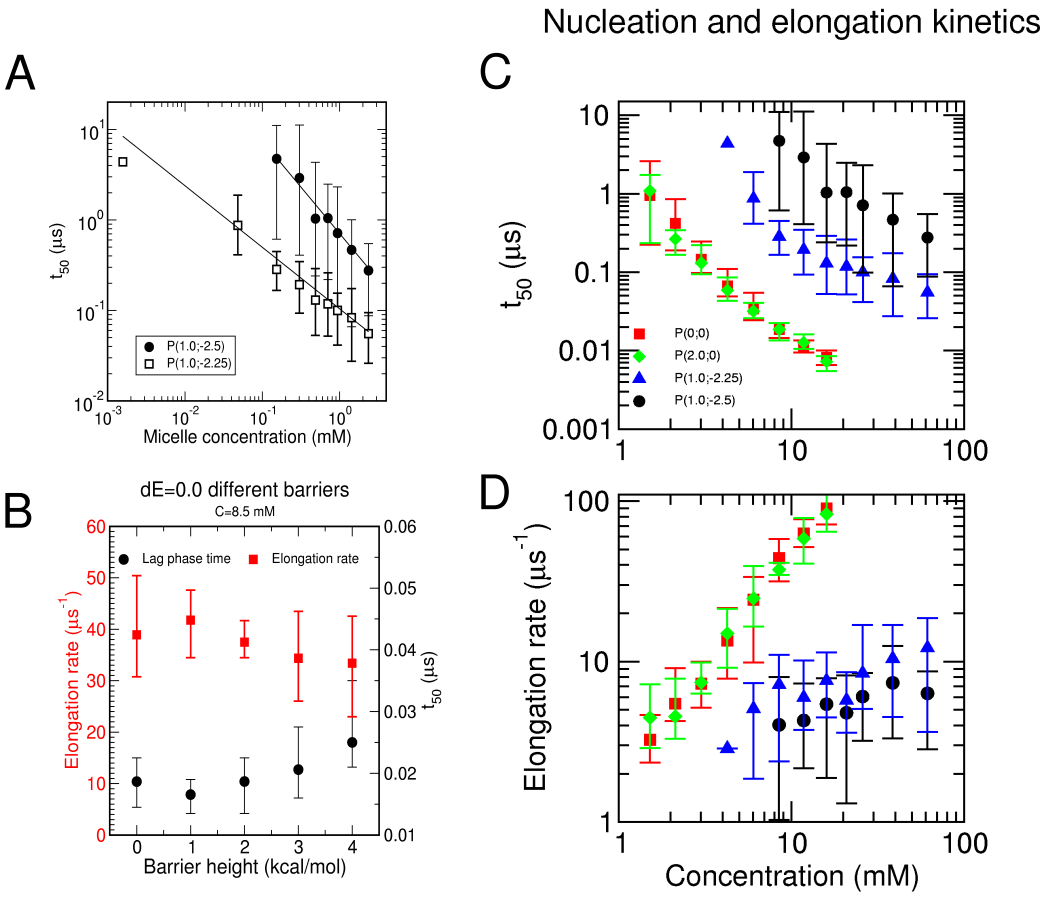


Figure 14: (A) Influence of the micelle concentration on the delay time t_{50} for β-unstable potentials $P(1.0;-2.5)$ and $P(1.0;-2.25)$. The power law fits are reported as continuous lines. The isolated data point at the lowest micelle concentration was not used for fitting for $P(1.0;-2.25)$. The error bars represent the minimum and the maximum value. (B) Effect on the delay time t_{50} (black circles) and the elongation rate k_e (red squares) of the barrier height variation for the potential $P(E_b;0)$. (C) Effect of concentration on delay time t_{50} for four potential models: $P(1.0;-2.5)$ black circles, $P(1.0;-2.25)$ blue triangles, $P(2.0;0)$ green diamonds, $P(0;0)$ red squares. The symbols represent the average value calculated on 15 simulations of $P(1.0;-2.5)$ and 10 simulations (all the others). The error bars represent the minimum and the maximum value. (D) Effect of concentration on the elongation rate k_e. Symbols and error bars as in (C).

C (mM)	P(0;0)	P(2;0)	P(1;-2.25)	P(1;-2.5)
1.52	10/10 (3.0 μs)	10/10 (3.0 μs)	n.a.	n.a.
2.13	10/10 (3.0 μs)	10/10 (3.0 μs)	n.a.	n.a.
3.01	10/10 (1.5 μs)	10/10 (1.5 μs)	n.a.	n.a.
4.26	10/10 (1.5 μs)	10/10 (1.5 μs)	1/10 (6.0 μs)	0/15 (3.0 μs)
6.04	10/10 (1.5 μs)	10/10 (1.5 μs)	10/10 (6.0 μs)	0/15 (6.0 μs)
8.50	10/10 (1.5 μs)	10/10 (1.5 μs)	10/10 (6.0 μs)	10/15 (12.0 μs)
11.8	10/10 (1.5 μs)	10/10 (1.5 μs)	10/10 (6.0 μs)	15/15 (12.0 μs)
16.0	10/10 (1.5 μs)	10/10 (1.5 μs)	10/10 (6.0 μs)	15/15 (12.0 μs)
20.9	n.a.	n.a.	10/10 (6.0 μs)	15/15 (3.0 μs)
25.9	n.a.	n.a.	10/10 (6.0 μs)	15/15 (3.0 μs)
38.7	n.a.	n.a.	10/10 (6.0 μs)	15/15 (3.0 μs)
61.5	n.a.	n.a.	10/10 (6.0 μs)	10/10 (3.0 μs)

Table 5: Table of all performed simulations for the concentration analysis of figure 14.C-D. The ratios indicate the number of nucleating trajectory over the number of independent simulations. The time reported in the brackets is the simulated time. (n.a.) the simulations were not performed at this concentration.

n_p time series of 15 independent simulations for the potential $P(1;-2.5)$, and 10 simulations for each of the remaining potentials. The delay time t_{50} is evaluated from the n_p time series as described in Section 3.1. Some of the $P(1;-2.5)$ runs were prolonged up to 12 μs (30 days on an Athlon 2800 GHz) to increase the number of nucleation events (Table 5).

The concentration dependence of the delay time and the elongation rate can be fitted by a power law $t_{50} = A_{50}C^{\gamma_{50}}$ and $k_e = A_eC^{\gamma_e}$, respectively, where C is the total monomer concentration. Results of the fit are reported in Table 6. Interestingly, the dependence of the rate of elongation on the concentration decreases significantly by increasing the stability of the protected state π. The reduced concentration dependence originates from competitive polymerizations, i.e., the elongation of the fibril and the presence of micelles. Furthermore, the concentration dependence of the delay time ($\gamma_{50} \neq 0$) for β-unstable potentials indicates that micelles promote the nucleation.

The nucleus sizes N^* can be extracted from the parameter γ_{50}, being $N^* = -2\gamma_{50}$, provided that *(a) the monomer concentration changes only by addition to and subtraction from polymers longer than the seed, (b) polymer formation by*

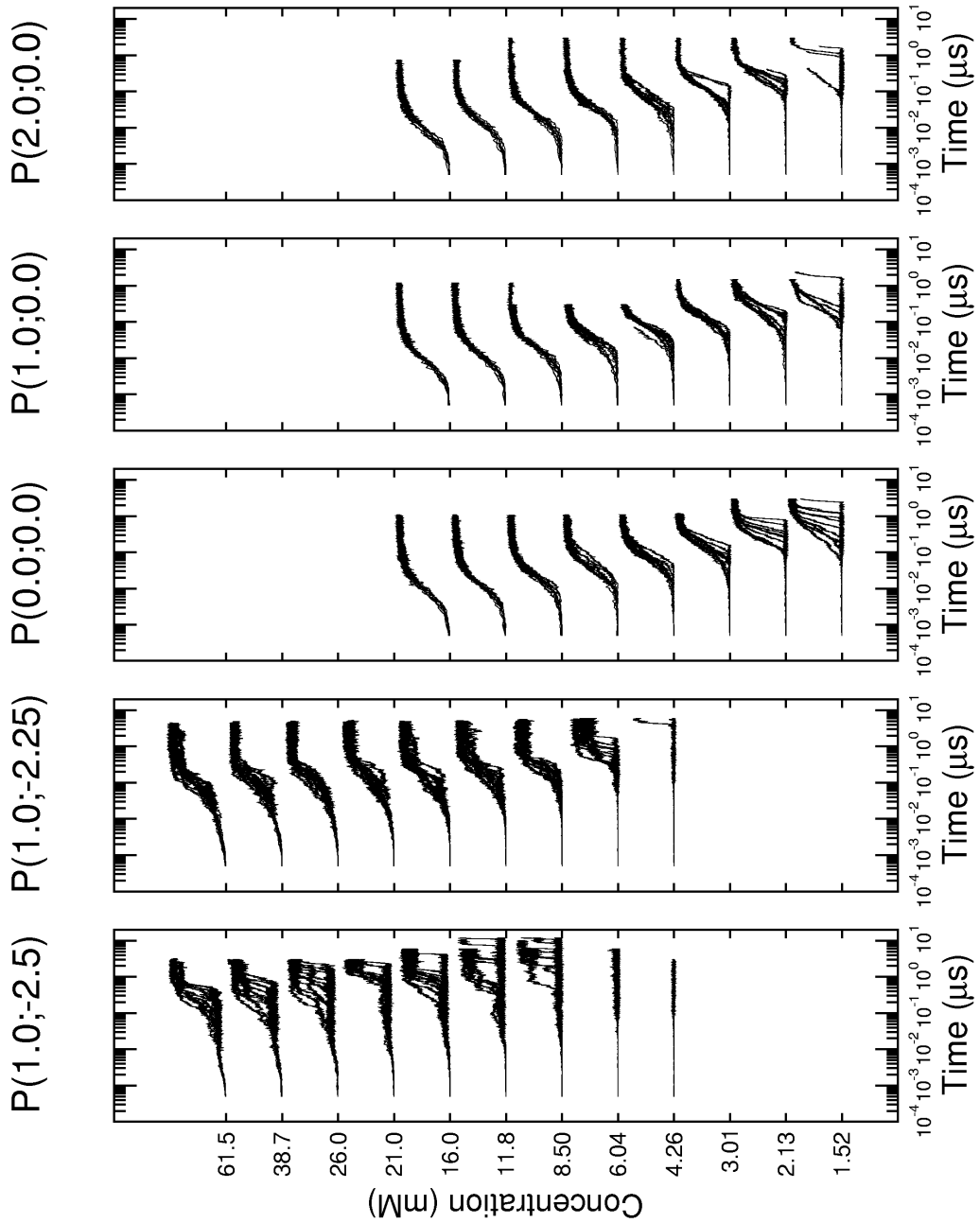


Figure 15: Concentration dependence for the time series of the number of parallel polar contacts n_p normalized to the maximum value. Time series belonging to the same concentration are reported in the same row. The five plots correspond to five different potential models.

Potential	A_{50}	γ_{50}	A_e	γ_e	A^*_{50}	γ^*_{50}
P(0;0)	2.73	-2.52	2.34	1.32	-	-
P(2;0)	1.75	-2.40	2.32	1.29	-	-
P(1;-2.25)	1.93	-0.916	2.24	0.409	0.104	-0.680
P(1;-2.5)	225.8	-1.80	2.45	0.258	0.708	-1.02

Table 6: Resulting fit parameters of the power law regression for the concentration dependence of kinetic observables. The delay time t_{50} and the elongation rate k_e were fitted for C greater that 8.5 mM for the $P(1.0; -2.25)$ potential (see Figure 14.C-D). The values of A^*_{50} and γ^*_{50} were obtained by fitting to the micelle concentration (Figure 14.A).

seed production is irreversible, and (c) the seed precursor is in pre-equilibrium with monomers [22]. These three assumptions are valid for β-stable models where only fibril and monomer species are produced, as demonstrated in Section 3.3. For $P(0;0)$ the nucleus size N^* is about 5, a value close to the one calculated with the probability of fibril formation (see Section 3.11 and Figure 6 of the main text). β-stable models, involving a small nucleus size, share the downhill polymerization mechanism described for the partially denaturated transthyretin [17]. On the other hand, the β-unstable models show a strong cohexistence of micellar and fibrillar oligomers in the lag and the elongation phases (see Section 3.5), therefore the hypothesis (a) cannot be fulfilled. Assuming that the nucleation process is first order to the micelle concentration, one can fit the delay time with a power law $t_{50} = A^*_{50} C_M^{\gamma^*_{50}}$, where C_M is the micelle concentration (see Figure 14.A and Section 3.4 for micelle concentration evaluation). The nucleus size $N^* = -2\gamma^*_{50}$, expressed in micelle units, is 1.36 for $P(1.0; -2.25)$ and 2.04 for $P(1.0; -2.5)$. Given the average aggregation number per micelle of 17.5 at $C = 8.5$ mM (see Figure 10) a value of 23.8 monomers and 35.7 monomers involved in the nucleation is obtained for $P(1.0; -2.25)$ and $P(1.0; -2.5)$, respectively. Strikingly, very similar values are obtained using the probability of fibril formation (see Figure 6 of the main text).

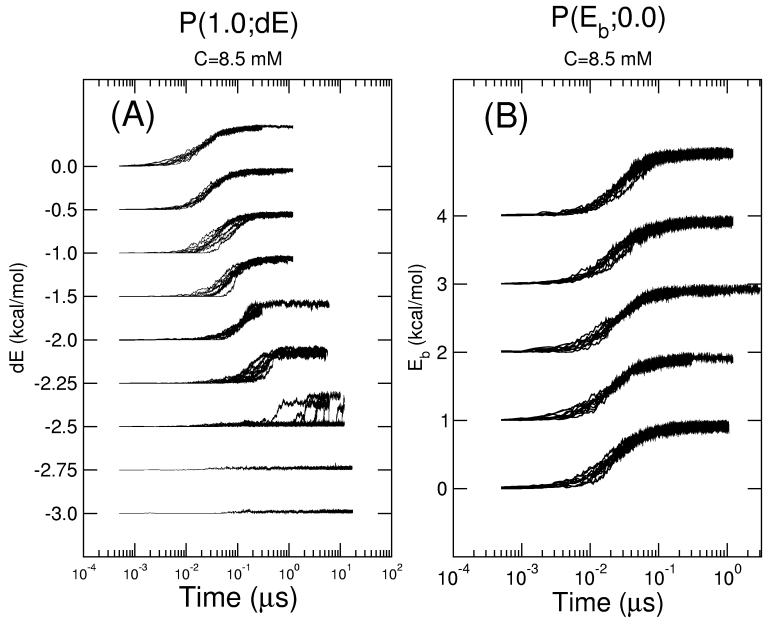


Figure 16: Stability, i.e., π-β free energy difference (A) and barrier (B) influence on the time series of the number of parallel polar contacts n_p normalized to the maximum value which corresponds to fibril. In plot (A) each row corresponds to a different value of dE (with constant E_b of 1.0 kcal/mol) while in plot (B) each row corresponds to a different value of E_b with constant $dE = 0$ kcal/mol.

3.8 Monomer energy landscape influence on the kinetics of fibril formation

To monitor the effects of the monomer energy surface a series of runs were performed at the concentration of 8.5mM. The left plot of Figure 16 displays the change of kinetics upon variation of the stability of the β-state without changing the $\beta \rightarrow \pi$ barrier E_b. The resulting rates and lag phase times are reported in Figure 2 of the main text. The right plot of Figure 16 shows the effects of variation of the barrier, keeping constant the stability of the β-state (see also Figure 14.B). No appreciable trend for the elongation and nucleation kinetics is observed for different values of E_b. These simulation results indicate that the aggregation kinetics of the model are mainly influenced by the relative stability of the β-aggregation prone state with negligible contribution of the $\beta \rightarrow \pi$ barrier.

3.9 Pathways of oligomeric aggregation

Since individual oligomers change their composition of monomers (i.e., in a given time interval an oligomer can absorb or release monomers to the solvent), it is crucial to define criteria for identifying oligomers along the simulation. Given the converged contact matrix at time t, $d^{(\infty)}(t)$, all oligomers at time t can be labeled: $A_1^t, A_2^t, ..., A_{n_t}^t$, where n_t is the number of oligomers. Each oligomer A_k^t has a size N_k^t and contains a list of tagged monomers $m_k^{1,t}, m_k^{2,t}, ..., m_k^{N_k^t,t}$. The time evolution of a single oligomer A_k^t at time $t+\tau$ is evaluated by comparing the monomer composition of every single oligomer present at time $t+\tau$. The similarity between two oligomers is defined as

$$S(A_k^t, A_l^{t+\tau}) = \sum_{i=1,N_k^t} \sum_{j=1,N_l^{t+\tau}} \delta(m_k^{i,t}, m_l^{j,t+\tau}) \tag{16}$$

where δ is the Kronecker function which is 1 if the compared monomers are the same. The time evolution $A_{k'}^{t+\tau}$ of oligomer A_k^t is defined as the oligomer with highest similarity:

$$S(A_k^t, A_{k'}^{t+\tau}) = \max_{l=1,n_{t+\tau}} (S(A_k^t, A_l^{t+\tau})) \tag{17}$$

If two or more $A_{k'}^{t+\tau}$ fulfill this equation, then the first labeled oligomer is chosen. A_k^t is then forwardly linked to $A_{k'}^{t+\tau}$, or equivalently $A_{k'}^{t+\tau}$ is assigned to the temporal successor of A_k^t. It is worth noting that:

1) *the temporal successor of an oligomer at time t is the oligomer at time* $t+\tau$ *that shares the highest number of monomers;*

2) *each oligomer can have a single temporal successor;*

3) *many oligomers can be forwardly linked to the same successor.*

By iterating this procedure, one can build the pathway of individual oligomers. Thus, the simulation trajectory is mapped to a network of temporally linked nodes (oligomers).

One natural definition for τ is the time difference between frames of the sim-

ulation, in our case $\tau = 0.5ns$, so that τ is smaller than the average life time of oligomers. The life time of an oligomer is the time needed to completely recycle monomers or to dissolve the oligomer. If this requirement for the time τ is not fulfilled the similarity can be ill-defined. Fibrils have a life time that is by far larger than τ being in the microsecond timescale (see Section 4.1). Metastable oligomers such as disordered aggregates or micelles have an estimated life time in the ten-nanoseconds scale. Unstable oligomers created by occasional collision of monomers have a life time slightly larger than τ.

3.10 β-subdomains time evolution and nucleus definition.

An important issue is the quantitative characterization of the nucleus. In literature, a nucleus is often defined as *the smallest marginally stable structured aggregate* [6]. In the framework of MD simulations a useful definition would be *the oligomer that has the same probability to either progress to a fibril or regress to the disordered state*. An analogous definition was applied to the folding transition state ensemble of two-state folders [23]. Using only the aggregation number N to distinguish the oligomers, this definition is problematic since an oligomer of given size can have a high morphological heterogeneity. It can contain no ordered aggregates, one subdomain ordered, two disjoint subdomains, and even more complicated features. A β-subdomain is a portion of the oligomer made of interacting β-monomers and the surrounding π-monomers can be considered as a local perturbation on the β-domains. With the previous nucleus definition it is possible to follow the dynamic evolution of β-subdomains in the context of their constituting oligomers. As a consequence the similarity procedure explained in Section 3.9 is applied to β-monomers only to identify the pathways of β-subdomains.

We define as progress variable of the ordered polymerization process, the aggregation number of the β-subdomain:

$$N(A_k^t) = N_k^t \tag{18}$$

where now A_k^t is an oligomer containing only β-monomers. A single trajectory is a collection of many independent pathways for the labeled β-subdomains (Figure 17). The pathway (A) is an unproductive event; first a β region appears in a π-only oligomer, it persists for a period of time, and disappears. (B) is a productive event; the appeared β-region spreads irreversibly into a fibril. Oligomers and their β-subdomains can interact in many different ways. Pathway (C) is a merging event while in (D) a β-region splitting is depicted. Pathway (E) is a combination of merging and division, while (F) is an interaction between oligomers that does not involve their β-subdomains.

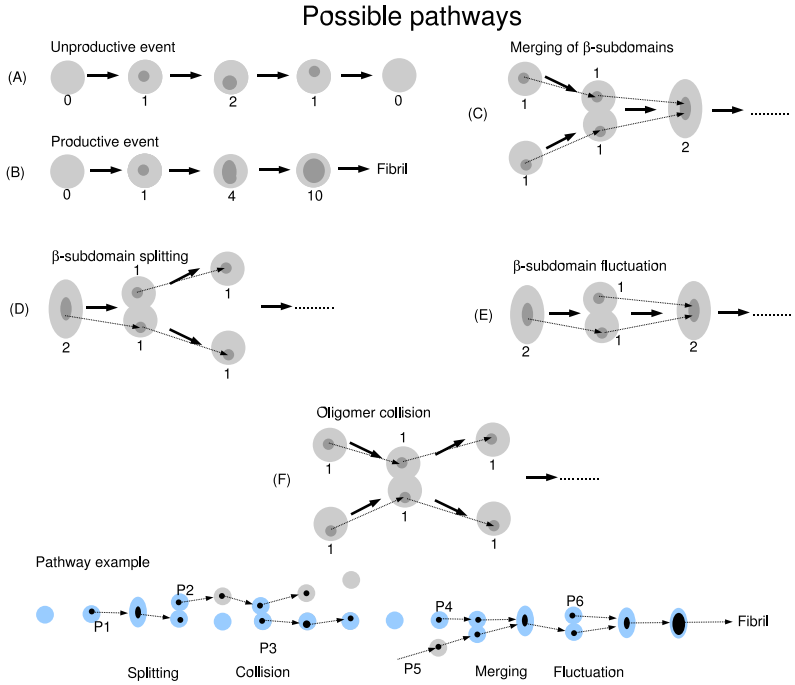


Figure 17: Classification of different pathways. The light gray regions schematize an oligomer, while the dark gray regions are the included β-subdomains. Bold arrows indicate the time evolution of entire oligomers, while the thin arrows are the time evolution of β-subdomains as defined by the similarity procedure (Equations 16 and 17). In the pathway example at the bottom, the dynamic evolution of an entire cluster (blue circles) and its β-subdomains (black dots) is depicted. Gray shaded regions are clusters interacting with the blue one.

3.11 Pathways analysis and probability of fibril formation

It is convenient to include an abstract state A_0, whose aggregation number is zero, to describe the beginning or the end of β-subdomain pathways. Another important state is the fibril state A_F; it is defined as the β-subdomain with aggregation number greater than 60. Given A_0 and A_F, an unproductive pathway is the trajectory of a β-subdomain that starts from A_0, and returns back to A_0, while a productive pathway is the trajectory of a β-subdomain that starts from A_0 and ends with the fibril state A_F. In the example of Figure 17, $P1$, $P2$ and $P3$ are unproductive while $P4$, $P5$ and $P6$ are productive. Given a set of trajectories, one can collect all productive and unproductive pathways P_i and define the probability of fibril formation of a β-subdomain A_{N_β} of size N_β as

$$p_{Ff}(N_\beta) = \frac{1}{M(A_{N_\beta})} \sum_{P_i \ni A_{N_\beta}} F(P_i) \tag{19}$$

where $M(A_{N_\beta})$ is the number of times that an aggregate of size N_β occurred in the simulations set. The sum runs over all P_is pathways that contain an aggregate A_{N_β}, and $F(P_i)$ is equal to 1 if the pathway P_i is productive, and is 0 otherwise. In this way the nucleus is unequivocally defined as *the oligomer containing a β-subdomain of size N^*_β with a probability of fibril formation $p_{Ff}(N^*_\beta)$ equal to* 0.5.

To finally characterize the nucleus it is useful to calculate its total aggregation number N^*. The average number of β-monomers as a function of the oligomer size N can be obtained by taking the exponential function on both side of Equation 15:

$$\frac{N_\beta(N)}{N_\pi(N)} = \exp\left[\frac{\Delta G_{\beta\pi}(N)}{kT}\right] \tag{20}$$

Using the fact that $N_\pi(N) = N - N_\beta(N)$ one has

$$N_\beta(N) = \frac{N \exp\left[kT^{-1}\Delta G_{\beta\pi}(N)\right]}{1 + \exp\left[kT^{-1}\Delta G_{\beta\pi}(N)\right]} \tag{21}$$

the average aggregation number N of an oligomer containing a β-subdomain with size N_β is obtained by numerical inversion of the function $N_\beta(N)$, $N = N(N_\beta)$ (Figure 6 of the main text), and the nucleus aggregation number is $N^* = N(N_\beta^*)$.

4 Supplementary analysis

4.1 Molecular Recycling

A molecular recycling mechanism has been observed by a combination of NMR spectroscopy and mass spectroscopy for an amyloid fibril formed from an SH3 domain [24]. To evaluate the recycling time of the coarse-grained model, simulations of mature fibrils in equilibrium with dispersed monomers are analyzed for the potential $P(1.0; -2.5)$ at all concentration values. The number of the unrecycled monomers $N_u(t)$ is defined as follows. First, all monomers belonging to the fibril at time $t = 0$ are labeled and counted. Then, at all times $t > 0$ the monomers that never detached from the fibril are counted and the resulting number is $N_u(t)$. In two of nine simulations, $N_u(t)$ goes to zero within 4 μs, which shows that all monomers initially belonging to the fibril have been recycled (Figure 18).

The number of unrecycled monomers $N_u(t)$ can be fitted with an exponential function:

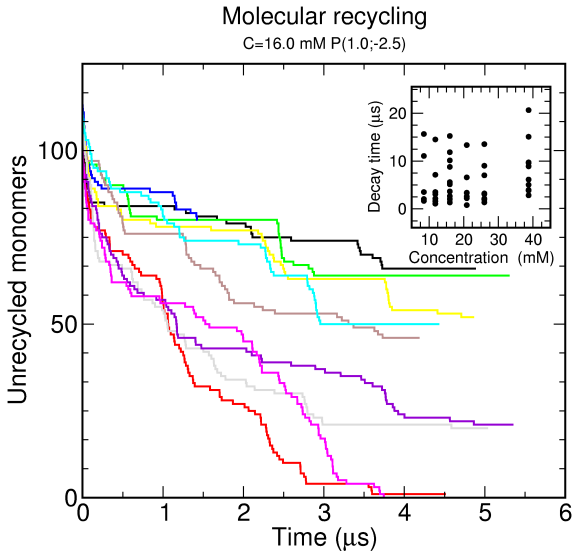


Figure 18: Number of unrecycled monomers N_u as a function of time for nine simulations started from a preformed equilibrated fibril. Simulations were performed at total concentration $C = 16.0$ mM and for the potential $P(1.0; -2.5)$. The values of decay time τ at different concentrations are reported in the inset.

$$N_u(t) = N_u(0)e^{-t/\tau} \tag{22}$$

where $N_u(0)$ is the initial value of monomers belonging to the fibril and τ is the time for the decay. In the inset of Figure 18, τ values for all concentrations and all independent simulations are reported. The decay times do not depend on the total concentration at which the fibril was formed.

4.2 Seeding

Fibril formation generally occurs via nucleation-dependent oligomerization with a lag time required for nucleus formation. This lag time can be abolished by using a seed, i.e., a small preformed fibril. To further evaluate the coarse-grained model and to validate the nucleus definition of Section 3.10, a seeding experiment is performed in *silico*. For the potential $P(1.0; -2.5)$ the smallest oligomer with high

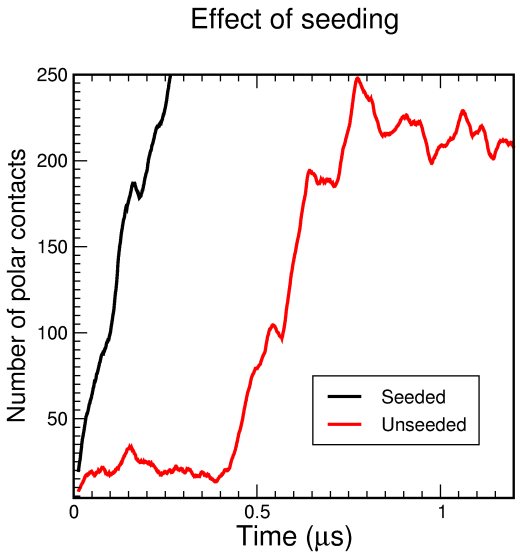


Figure 19: Number of polar contacts n_p along the time for a spontaneous (red line) and seeded trajectory both at 8.5 mM and $P(1.0; -2.5)$ potential. The unseeded simulation shown here is the fastest nucleation observed at this concentration while nucleation in the other runs are about one order of magnitude slower (see Figure 16 left). The spontaneous trajectory reaches a plateau at about 225 polar contacts because it was run with 125 monomers, whereas the seeded trajectory was run with a total of 1000 monomers.

probability of fibril formation is isolated from a trajectory at 38.7 mM. This post-critical oligomer consists of 60 monomers and has a probability of fibril formation of 98% according to the definition of Section 3.10 and to Figure 4 of the main text. The oligomer is introduced in a box with 940 dispersed monomers at a total concentration of 8.5 mM. This is the lowest concentration that displayed a nucleation for this potential (see Figure 15 and Table 5). The average lag phase for the spontaneous nucleation is around 5 μs (see Figure 14.C) and the minimal lag phase time observed is 0.4 μs (Figure 19). Strikingly, the lag phase is completely abolished in the seeded simulation (Figures 19 and 20).

On the other hand for the $P(0;0)$ potential seeding does not influence the kinetics (data not shown), as expected for downhill polymerization [17].

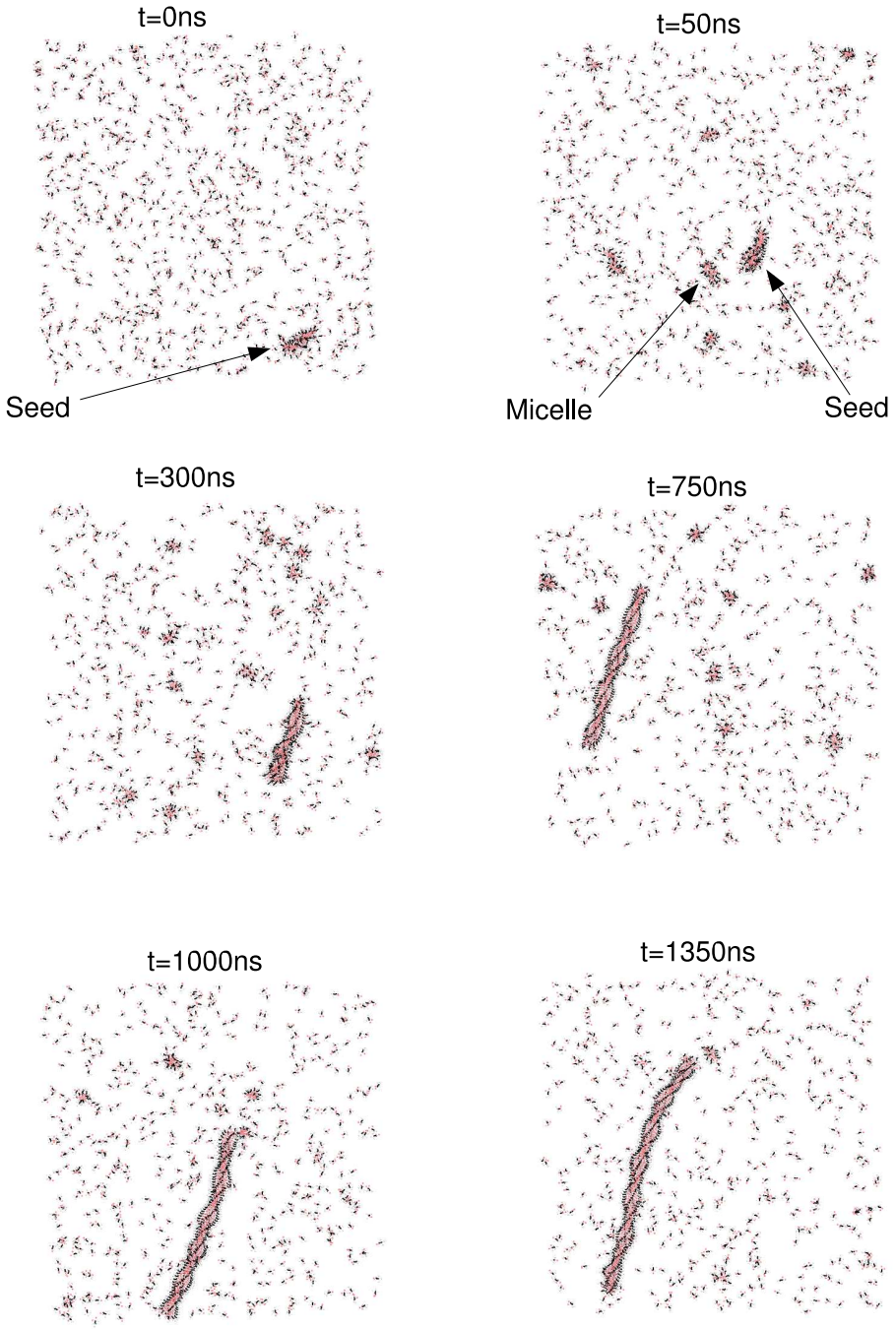


Figure 20: Six illustrative snapshots of the seeding trajectory. At the start ($t = 0$ ns) there are 60 monomers in the post-critical oligomer (i.e. the seed) and 940 monodispersed monomers. The total concentration is 8.5 mM. Within the first 50 ns micelles are nucleated, and progressively disappear during the fibril elongation phase.

References

1. MacKerell, A. D. J., Feig, M. & Brooks, C. L. r. Improved treatment of the protein backbone in empirical force fields. *J Am Chem Soc* **126**, 698–699 (2004).

2. Hansen, J.-P. & McDonald, I. R. *Theory of simple liquids* (Academic Press, San Diego, 1996).

3. Soreghan, B., Kosmoski, J. & Glabe, C. Surfactant properties of Alzheimer's A beta peptides and the mechanism of amyloid aggregation. *J Biol Chem* **269**, 28551–28554 (1994).

4. Serio, T. R. *et al.* Nucleated conformational conversion and the replication of conformational information by a prion determinant. *Science* **289**, 1317–1321 (2000).

5. Yong, W. *et al.* Structure determination of micelle-like intermediates in amyloid beta -protein fibril assembly by using small angle neutron scattering. *Proc Natl Acad Sci U S A* **99**, 150–154 (2002).

6. Rhoades, E. & Gafni, A. Micelle formation by a fragment of human islet amyloid polypeptide. *Biophys J* **84**, 3480–3487 (2003).

7. Bitan, G. *et al.* Amyloid beta -protein (Abeta) assembly: Abeta 40 and Abeta 42 oligomerize through distinct pathways. *Proc Natl Acad Sci U S A* **100**, 330–335 (2003).

8. Sabatè, R. & Estelrich, J. Evidence of the existence of micelles in the fibrillogenesis of β-amyloid peptide. *J Phys Chem* **109**, 11027–11032 (2005).

9. Ryckaert, J. P., Ciccotti, G. & Berendsen, H. J. C. Numerical integration of the Cartesian equation of motion of a system with constraints: Molecular dynamics of n-alkanes. *J. Comp. Phys.* **23**, 327–341 (1977).

10. Brooks, B. R. *et al.* CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **4**, 187–217 (1983).

11. Aggeli, A. *et al.* Hierarchical self-assembly of chiral rod-like molecules as a model for peptide beta -sheet tapes, ribbons, fibrils, and fibers. *Proc Natl Acad Sci U S A* **98**, 11857–11862 (2001).

12. Goldsbury, C. S. *et al.* Polymorphic fibrillar assembly of human amylin. *J Struct Biol* **119**, 17–27 (1997).

13. Petkova, A. T. *et al.* Self-propagating, molecular-level polymorphism in Alzheimer's beta-amyloid fibrils. *Science* **307**, 262–265 (2005).

14. Chiti, F. *et al.* Mutational analysis of the propensity for amyloid formation by a globular protein. *EMBO J* **19**, 1441–1449 (2000).

15. Chiti, F. *et al.* Solution conditions can promote formation of either amyloid protofilaments or mature fibrils from the HypF N-terminal domain. *Protein Sci* **10**, 2541–2547 (2001).

16. Dobson, C. M. Protein misfolding, evolution and disease. *Trends Biochem Sci* **24**, 329–332 (1999).

17. Hurshman, A. R., White, J. T., Powers, E. T. & Kelly, J. W. Transthyretin aggregation under partially denaturing conditions is a downhill polymerization. *Biochemistry* **43**, 7365–7381 (2004).

18. Johnson, S. C. Hierarchical clustering schemes. *Psychometrika* **32**, 241–254 (1967).

19. O'Nuallain, B., Shivaprasad, S., Kheterpal, I. & Wetzel, R. Thermodynamics of A beta(1-40) amyloid fibril elongation. *Biochemistry* **44**, 12709–12718 (2005).

20. Esler, W. P. *et al.* Alzheimer's disease amyloid propagation by a template-dependent dock-lock mechanism. *Biochemistry* **39**, 6288–6295 (2000).

21. Collins, S. R., Douglass, A., Vale, R. D. & Weissman, J. S. Mechanism of prion propagation: amyloid growth occurs by monomer addition. *PLoS Biol* **2**, e321 (2004).

22. Goldstein, R. F. & Stryer, L. Cooperative polymerization reactions. Analytical approximations, numerical examples, and experimental strategy. *Biophys J* **50**, 583–599 (1986).

23. Du, R., Pande, V. S., Grosberg, A. Y., Tanaka, T. & Shakhnovich, E. I. On the transition coordinate for protein folding. *J. Chem. Phys.* **108**, 334–350 (1998).

24. Carulla, N. *et al.* Molecular recycling within amyloid fibrils. *Nature* **436**, 554–558 (2005).

6.3 Pathways and intermediates of amyloid fibril formation. [J. Mol. Biol. 2007, *374*, 917].

doi:10.1016/j.jmb.2007.09.090 J. Mol. Biol. (2007) 374, 917–924

JMB

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

COMMUNICATION

Pathways and Intermediates of Amyloid Fibril Formation

Riccardo Pellarin, Enrico Guarnera and Amedeo Caflisch*

Department of Biochemistry
University of Zürich
Winterthurerstrasse 190
CH-8057 Zürich, Switzerland

Received 18 July 2007;
received in revised form
13 September 2007;
accepted 28 September 2007
Available online
4 October 2007

The lack of understanding of amyloid fibril formation at the molecular level is a major obstacle in devising strategies to interfere with the pathologies linked to peptide or protein aggregation. In particular, little is known on the role of intermediates and fibril elongation pathways as well as their dependence on the intrinsic tendency of a polypeptide chain to self-assembly by β-sheet formation (β-aggregation propensity). Here, coarse-grained simulations of an amphipathic polypeptide show that a decrease in the β-aggregation propensity results in a larger heterogeneity of elongation pathways, despite the essentially identical structure of the final fibril. Protofibrillar intermediates that are thinner, shorter and less structured than the final fibril accumulate along some of these pathways. Moreover, the templated formation of an additional protofilament on the lateral surface of a protofibril is sometimes observed as a collective transition. Conversely, for a polypeptide model with a high β-aggregation propensity, elongation proceeds without protofibrillar intermediates. Therefore, changes in intrinsic β-aggregation propensity modulate the relative accessibility of parallel routes of aggregation.

Keywords: amyloid protofibrils; fibril growth; aggregation pathways; molecular dynamics simulations; Alzheimer's disease

Edited by F. E. Cohen

The link between protein aggregates and progressive neurodegenerative pathologies, like Alzheimer's, Parkinson's, Huntington's and prion diseases, exists but is not clear.[1,2] Despite the medical relevance of these devastating diseases, little is known about the aggregation process itself and, most importantly, how to safely inhibit the formation of toxic species. Experimental evidence indicates that early aggregates, e.g. soluble oligomers and protofibrils, have a critical role in promoting pathological effects in amyloid disorders.[3,4] As an example, the E22G mutation of the Alzheimer's peptide (Aβ) enhances protofibril formation,[5] and plaque formation is more aggressive than for wild-type Aβ in transgenic mice.[6] Also, mutations of α-synuclein that are related to early-onset forms of Parkinson's disease can produce protofibrils efficiently.[7] Yet, the molecular details and the mechanisms leading to the toxicity of these prefibrillar aggregates are only partially understood. In fact, the transient character of oligomeric precursors hinders the complete understanding of their formation process and structural details.

*Corresponding author. E-mail address: caflisch@bioc.uzh.ch.

The available experimental evidence *in vitro* indicates that the kinetics of fibril formation are complex and can be often separated into a nucleation (or lag) phase and an elongation phase,[8] followed by the equilibrium between isolated polypeptides and the fibrils.[9] Multistep kinetics with the presence of intermediates have also been reported.[10] Pathways of fibril formation, fibril morphologies and stability of protofibrillar intermediates are influenced strongly by experimental conditions (e.g. protein concentration, pH and ionic strength),[11] and elongation rates can depend on the stability of aggregation prone folding intermediates.[12]

Theoretical models have been developed to investigate the amyloid aggregation mechanism[13–15] and predict the rates[16] but strong assumptions like the irreversible association of polypeptide chains onto the fibril[13,16] are not consistent with the interpretation of experimental results.[9,17] Computer simulations using low-resolution models, which employ a simplified representation of protein geometry and energetics, have provided insights into the basic physical principles underlying protein aggregation in general,[18–20] and ordered amyloid aggregation.[21–28] However, they do not explain the wide range of aggregation processes emerging from a variety of

biophysical studies.[11,29] Atomistic models have shed some light on oligomeric aggregates and the very early steps of fibril formation,[30–36] but all-atoms simulations aimed at reproducing the kinetics and investigating the pathways of fibril formation are computationally expensive and difficult to analyze.

Earlier, we developed a phenomenological coarse-grained model of an amphipathic polypeptide and used it for exploring the kinetics of nucleation and the rates of and elongation by Langevin dynamics simulations.[37] To allow for efficient sampling, the conformational landscape of the isolated monomer was simplified such that only two states are considered: the amyloid-competent (β) and the amyloid-protected (π) states (Figure 1). In the β-state, the parallel orientation of the two intramolecular dipoles favors ordered aggregates with intermolecular dipolar interactions parallel with the fibril axis. Conversely, the π-state represents the ensemble of all polypeptide conformations that are not compatible with self-assembly into a fibril. At physiological temperature the isolated monomer undergoes a reversible isomerization from the π-state to the β-state. The energy difference between these two states can be interpreted as the β-aggregation propensity of a polypeptide sequence. For instance when $dE = E_\pi - E_\beta = 0.0$ kcal/mol, the π and β states are equally populated, whereas for $dE = -1.5$ kcal/mol and -2.5 kcal/mol the π-state is about 15 and 100 times more populated than the β state, respectively. It was found that despite the essentially identical structure of the final fibril, ordered aggregation of a polypeptide with a stable β-state follows a pathway devoid of stable intermediates, while on-pathway micellar oligomers (with hydrophilic surface and hydrophobic interior) were observed during the nucleation phase of a polypeptide with a β-state that is marginally stable. In other words, high and low β-prone sequences show significantly different nucleation processes. These two models are termed β-stable and β-unstable, respectively, and the passage from one regime to the other was achieved by varying solely the parameter dE.[37] The focus of our previous study was on the nucleation phase, while the elongation mechanism and pathway(s) were not investigated. Here, for each of four

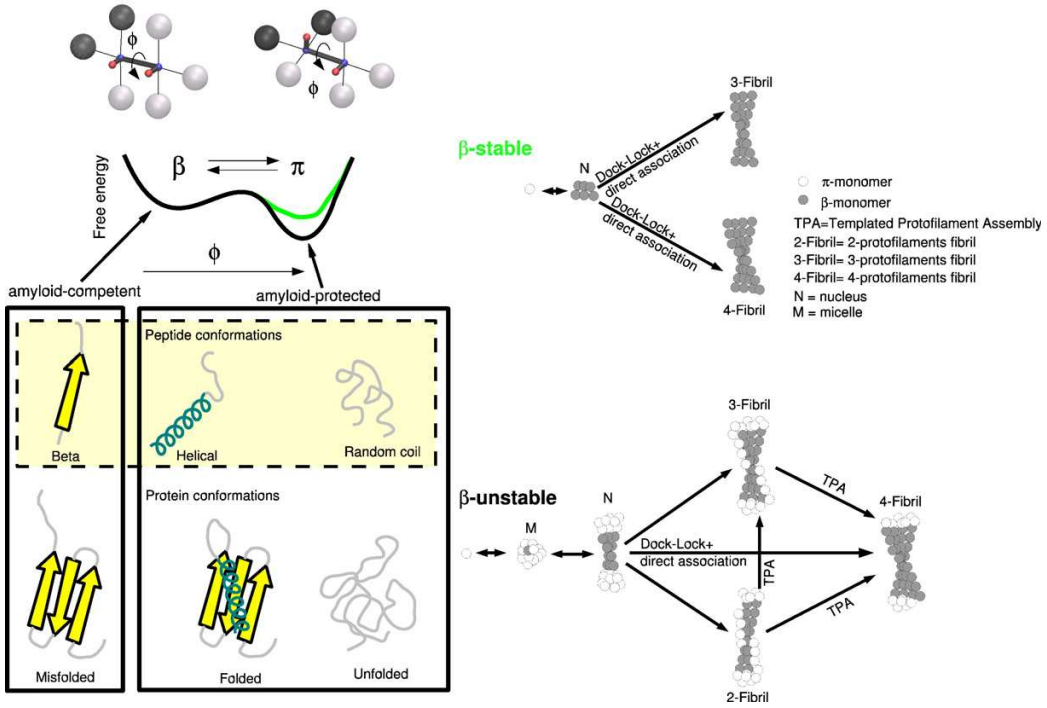


Figure 1. The model and aggregation pathways. Left: Sticks and beads representations of the monomer in the amyloid-competent state β and the amyloid-protected state π. The large spheres are hydrophobic (black) and hydrophilic (gray), while the two dipoles are shown with small red and blue spheres. The β and π states of the monomer are shown on top of the two corresponding minima of the free energy, plotted as a function of the dihedral angle ϕ of the two dipoles. Note that the population of monomers in the β-state decreases by lowering the free energy of the π-state, as indicated by the green and black profiles. For each value of the β-aggregation propensity dE ($dE = E_\pi - E_\beta = -1.5, -2.0, -2.25, -2.5$ kcal/mol) 100 Langevin dynamics runs with different initial assignments of the velocities were started from 125 monomers uniformly distributed in a box with random orientations. All simulations were carried out at a temperature of 310 K and a concentration of 8.5 mM with the same force-field parameters as those used previously.[37] Results discussed in this work refer mainly to the β-stable ($dE = -1.5$ kcal/mol) and the β-unstable ($dE = -2.5$ kcal/mol) models. Right: Observed aggregation pathways for the β-stable and β-unstable models. The elongation pathways of the latter are more heterogeneous than those of the former.

polypeptide models (four values of dE that range from β-stable to β-unstable) 100 Langevin dynamics runs were performed to explore the elongation phase; i.e. the pathway(s) leading from the nucleus to the final fibril.

The present work was motivated by the following two questions: what is the influence of the intrinsic β-aggregation propensity on the mechanism of fibril elongation? and are there multiple pathways and/or intermediates? From a detailed analysis of the simulations (started from 125 coarse-grained monomers in a monodisperse state), a rich scenario of alternative pathways, some with prefibrillar intermediates, emerges only for monomers with a low β-aggregation propensity. The simulation results go beyond the fibril formation mechanisms suggested on the basis of biophysical measurements, and have strong implications for the design of inhibitors of amyloid aggregation.

Terminology

A rigorous terminology for the early aggregates and intermediates of amyloid self-assembly observed *in vitro* has been recently summarized.[38,39] Because the computer simulations allow for the detailed investigation of individual oligomers as well as prefibrillar states and the final fibril, it is useful and straightforward to define the following nomenclature: a protofilament is a file of monomers with intermolecular dipolar interactions parallel with its axis; a protofibril is a transient structure that consists of two to three protofilaments with large unstructured regions; and the final fibril is a fully ordered aggregate of three to four protofilaments. In the model used here, the fibril is stabilized by intermolecular dipolar interactions within each protofilament and van der Waals interactions between hydrophobic beads.[37]

Aggregation state network

An aggregate consists of monomers whose mutual minimal distances are less than 6 Å, and it is isolated using a clustering procedure as described.[37] Three progress variables are used to monitor the aggregation process: the size of the largest aggregate N_{la}, the number of monomers in the β-state within the largest aggregate N_{la}^{β}, and the number of protofilaments in the largest aggregate N_{la}^{pf}. Note that the range of N_{la} is limited by the size of the simulated system ($1 \leq N_{la} \leq 125$). The number of protofilaments within a single aggregate is calculated by counting the files of monomers in the β-state with intermolecular dipolar interactions. Let N_f be the number of such files present into a given aggregate, and $\omega_1, \cdots, \omega_{Nf}$ the number of monomers in each file (with $\omega_i > 10$ to reduce noise). The number of protofilaments in aggregate a, N_a^{pf}, is thus defined as:

$$N_a^{pf} = \frac{\left(\sum_{i=1}^{N_f} \omega_i\right)^2}{\sum_{i=1}^{N_f} \omega_i^2} \qquad (1)$$

This definition prevents counting small isolated files whose formation is a result of thermal fluctuations, enhancing the signal to noise ratio with respect to N_f. Two limiting cases are useful to explain this variable. In the case that all files have the same size (i.e. $\omega_1 = \ldots = \omega_{Nf}$), the protofilament number N_a^{pf} is equal to the number of files N_f. In the case where a single ω_i predominates ($\omega_i \gg \omega_k$ for all k different from i) N_a^{pf} tends to 1. The number of protofilaments in the largest aggregate N_{la}^{pf} is thus the function N_a^{pf} applied to the largest of all aggregates present in the simulation volume. Selected time series of N_{la}, N_{la}^{β} and N_{la}^{pf} are reported in Figure 2.

The aggregation state network (Figure 3) is a graph in which states and direct transitions observed during the Langevin dynamics simulations are displayed as nodes and links, respectively.[40] Furthermore, the size of each node reflects the statistical weight of the corresponding state. In this way, metastable states and their dynamic connectivity are illustrated without requiring projections onto arbitrarily chosen reaction coordinates.[41] Micellar oligomers (white nodes, $N_{la} \sim 20, N_{la}^{pf} = 0$), which are spherical aggregates whose core consists of the hydrophobic spheres of the monomers (see inset A of Figure 3),[37] and fibrils (red nodes, $N_{la} \sim 100$, $N_{la}^{pf} = 4$) are the most populated states during the lag phase and the final equilibrium, respectively. Strikingly, a greater variety of aggregation mechanisms emerges for the β-unstable (Figure 3, bottom) than the β-stable polypeptide model (Figure 3, top). In particular, the former shows the presence of intermediates, i.e. protofibrils consisting of only two (green nodes) or three (blue nodes) protofilaments. Moreover, the aggregation state network qualitatively illustrates that the protofibrils are metastable and it displays broad transition regions between the two-protofilament state and the three-protofilament state, as well as between the latter and the final fibril.

Templated protofilament assembly

Previously, the elongation rate was found to increase according to the population of the amyloid-competent state,[37] but the underlying mechanism of elongation was not investigated. Using the Markov chain formalism (see the Supplementary Data) it is possible to estimate the rate of association of a monomer to a fibril followed by the isomerization from the amyloid-protected state to the amyloid-competent state (k_{fibril}). An alternative process is the monomer isomerization in the solvent followed by association ($k_{solvent}$). In their analytical model of fibril elongation, Massi and Straub have illustrated these two pathways as the route of monomer association to the fibril followed by isomerization (deposition and reorganization in Figure 4 of Massi & Straub[14]) and the route of direct association (direct deposition in Figure 5 of Massi & Straub[14]). The former pathway corresponds to the dock-lock mechanism.[42–44] Hence, the ratio $k_{fibril}/k_{solvent}$ measures the efficiency of the dock-lock mechanism; it is

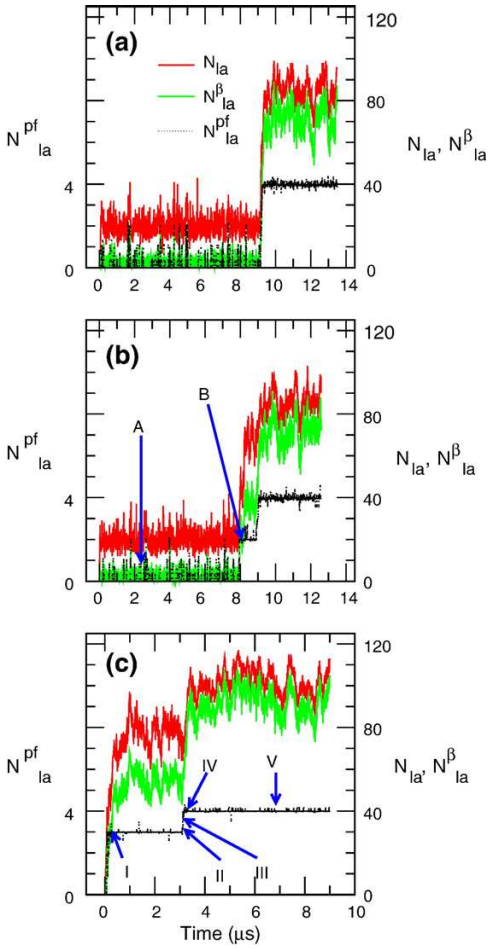


Figure 2. Protofibrillar intermediates and pathway heterogeneity. The time-series of three progress variables are used to monitor the evolution of the largest aggregate (1a) in the β-unstable simulations: The number of protofilaments N_{la}^{pf} (black curve with the *y*-axis description on the left; note that this quantity is evaluated by equation (1) and can be non-integer), the size of the largest aggregate N_{la} and the number of monomers in β-state N_{la}^{β} (red and green curves, respectively, with the *y*-axis description on the right). The three runs shown are representative of (a) elongation without intermediates, and (b) with two-filament or (c) three-filament protofibrillar intermediates. Templated protofilament assembly is observed at about 9 μs in (b) and at about 3 μs in (c), and the snapshots labeled are shown in Figure 3.

3.6 for the β-unstable model and 6.6 for the β-stable model. In both cases, the rate of conversion of a monomer bound to a fibril exceeds that in solution, suggesting that the elongation is dominated by a dock-lock mechanism. Nevertheless, this mechanism does not exclude collective conversions.

Representative time series of the number of protofilaments N_{la}^{pf} are shown by a black curve in Figure 2 for the β-unstable model. Metastable intermediates are observed in about half of the runs (see Supplementary Data). Interestingly, during some of the fast transitions from a three-protofilament aggregate to the final fibril (or sporadically from two to three-protofilament protofibrils) the size of the largest aggregate (red line) does not change significantly, whereas its number of monomers in the β-state (green line) increases abruptly, e.g. at about 9 μs and 3 μs in Figure 2(b) and (c), respectively. The collective conversion of monomers from the amyloid-protected to the amyloid-competent state is a consequence of the templated assembly of the fourth filament on the metastable protofibril consisting of three protofilaments (Figure 3 insets I–V). In other words, a file of monomers in the amyloid-protected conformation accumulates, first without forming intermolecular dipolar interactions, along the exposed hydrophobic surface of the three-protofilament aggregate (blue monomers in inset I). This event is then followed by a collective transition during which all monomers in the file convert to the β-state, which is stabilized by both intermolecular dipole interactions within the fourth protofilament and van der Waals interactions with monomers in the other three protofilaments (insets II–IV). The templated-assembly mechanism observed in the simulations is consistent with measurements of insulin aggregation by atomic force microscopy.[45] Moreover, protofibril maturation into fibrils is irreversible under the conditions used in the present simulations, i.e. 310 K and 8.5 mM (see Figure 2). Irreversibility has been suggested on the basis of the temporal increase in average protofibril size measured by quasi-elastic light-scattering spectroscopy.[46]

Analysis of the time series of the β-stable model does not reveal any event of templated protofilament formation. In fact, fibrils composed of three protofilaments contain as many monomers in the β-state as the mature four-protofilament fibril (see Figure 4(d)); thus, the formation of the fourth protofilament corresponds to a redistribution of monomers in the β-state among the protofilaments.

Size and structural characterization of protofibrils

The size distribution of the two and three-protofilament aggregates are different and depend on the β-aggregation propensity of the monomer (Figure 4). During the elongation phase, intermediates with two protofilaments are observed mainly for the β-unstable model (peak at $N_{la} \sim 70$). By raising the β-aggregation propensity (from $dE = -2.5$ kcal/mol to $dE = -1.5$ kcal/mol) there is a decrease in the average aggregation size of two-protofilament aggregates. Protofibrils consisting of three protofilaments are observed during the elongation phase of all models. Notably, by increasing the β-aggregation tendency, the number of runs with on-pathway intermediates decreases monotonically, which reflects the lower heterogeneity of pathways for the β-stable model.

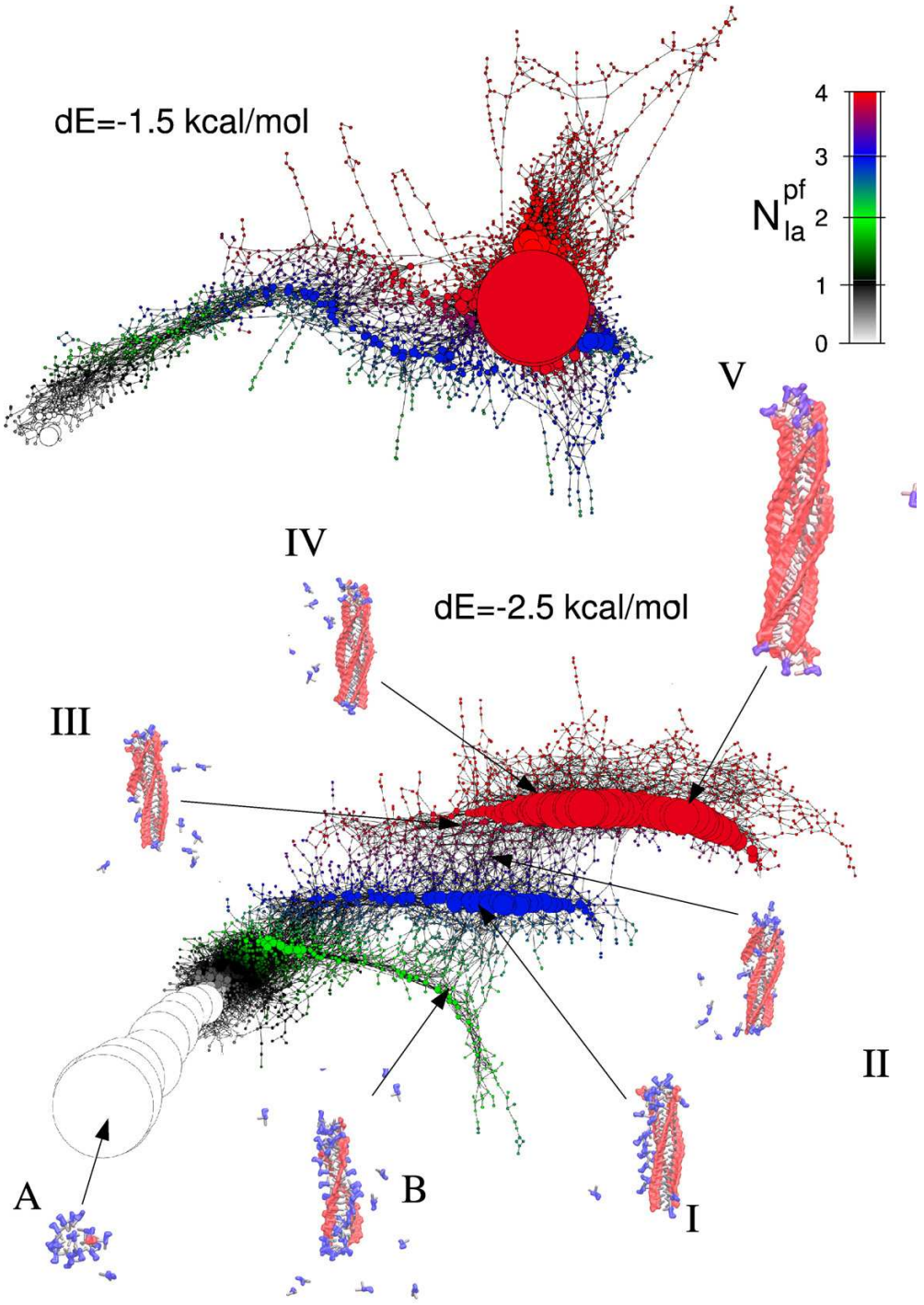


Figure 3. Aggregation state network. The size of the largest aggregate N_{la} and its number of protofilaments N_{la}^{pf} were used to cluster all snapshots into states (i.e. nodes of the network). The size and color of the nodes correspond to the statistical weight and the number of protofilaments N_{la}^{pf}, respectively. Links are direct transitions within 0.5 ns (10,000 steps of 50 fs each) of Langevin dynamics. All the states and the transitions that have been explored by the simulations are represented in these networks. Note the much higher heterogeneity of protofibrillar intermediates for the β-unstable (dE= −2.5 kcal/mol, bottom) than the β-stable (dE= −1.5 kcal/mol, top) model. The insets show the structures of the largest aggregates from the snapshots labeled in Figure 2. In these structures, monomers in the amyloid-competent conformer β and amyloid-protected conformer π are in red and blue, respectively. Furthermore, hydrophobic spheres are gray and hydrophilic spheres are not shown for visual clarity.

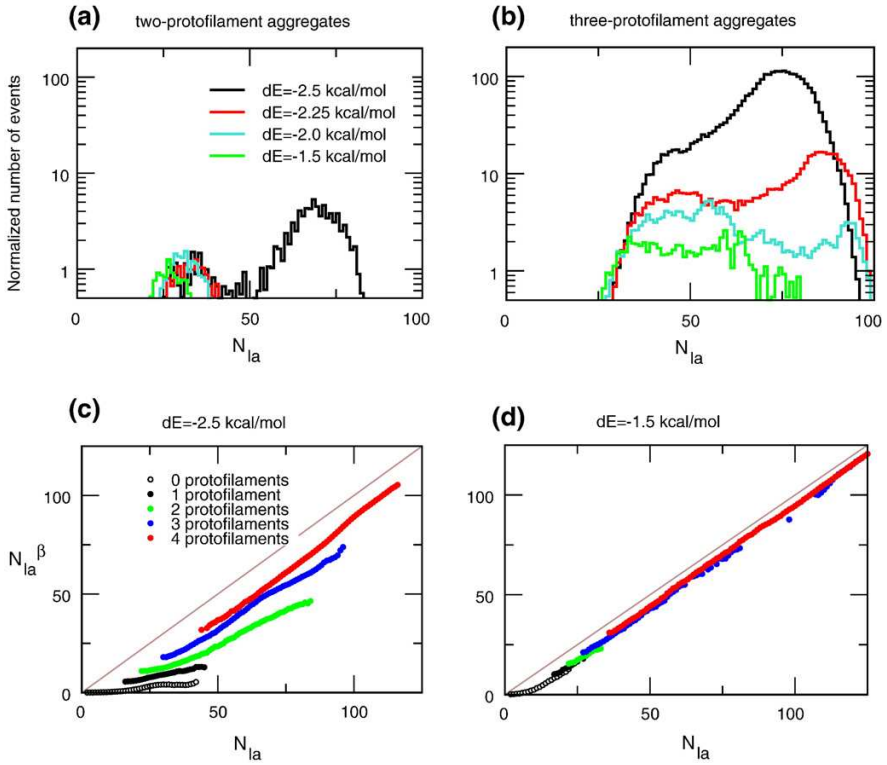


Figure 4. Size distribution of (a) two-protofilament and (b) three-protofilament protofibrils during fibril growth. The histograms are built by counting the trajectory frames in which the largest aggregate contains either two or three protofilaments. The frames are collected only during the elongation phase, i.e. after the nucleation step and before reaching the final monomer/fibril equilibrium. Average value of the number of monomers in β-state contained into the largest aggregate, as a function of the size of the largest aggregate for the (c) β-unstable and (d) the β-stable models.

For the β-unstable model protofibrils are thinner, shorter and more disordered than the final fibril. The protofibrils and fibrils of this model often present deposits of monomers in the π-state that are not involved in intermolecular dipole interactions and are highly disordered (blue monomers in the insets of Figure 3). The ratio between the number of monomers in the β-state and the total number of monomers N_{la}^{β}/N_{la} is significantly smaller than 1, even for fibrils consisting of four protofilaments (Figure 4(c)). The deviation is due mainly to the fibril ends that are populated by monomers in the π-state (see Figure 3 inset V). Furthermore, protofibrils with two or three protofilaments contain less monomers in the β-state than the four-protofilament fibril of the same size. Conversely, for the β-stable model the N_{la}^{β}/N_{la} ratio is always close to 1, and aggregates of three protofilaments can have more than 100 monomers (Figure 4(d)).

Conclusions

The self-assembly process of an amphipathic polypeptide has been investigated by multiple Langevin dynamics simulations using a coarse-grained model whose simplicity allows for the sampling of hundreds of fibril formation events. By varying a single parameter of the model, namely the relative stability of the amyloid-competent and amyloid-protected states of the polypeptide (β-aggregation propensity), interesting insights into elongation pathways and protofibrillar intermediates have been obtained. Two main observations emerge from the simulation results.

First, the roughness of the free-energy surface governing the aggregation process and the heterogeneity of pathways of fibril elongation increase by reducing the β-aggregation propensity. Hence, a mutation that decreases the β-aggregation tendency could result in greater variety of prefibrillar aggregates. Interestingly, these simulation results provide a possible explanation for the enhanced *in vitro* formation of oligomers and protofibrils of the Arctic mutant (E22G) of the Alzheimer's Aβ peptide,[5] and the A30P mutant of α-synuclein.[7] In fact, among the 20 standard amino acids, glycine and proline residues have the weakest propensity of β-sheet formation,[47] and β-aggregation.[48]

Second, a mechanism of templated protofilament assembly is sometimes observed during fibril growth. Although the elongation is accomplished mainly by dock-lock monomer addition at the

growing ends, the formation of an ordered protofilament can occur at the lateral surface of a protofibril by collective interconversion of a file of previously deposited monomers. This mechanism is particularly frequent for the model with low β-aggregation propensity, where, due to the frustration of the conformational landscape, the isomerization of a single monomer is strongly disfavored.

In conclusion, the simulation results provide strong evidence of multiple routes of polypeptide self-assembly. Notably, a reduction of the intrinsic β-aggregation propensity induces higher pathway heterogeneity and on-pathway protofibrillar intermediates. Given the experimental evidence of toxicity of prefibrillar aggregates, one is tempted to speculate that therapeutic strategies aimed at reducing fibril-formation propensity (e.g. stabilization of the folded state by small molecules) might paradoxically promote the accumulation of toxic species.

Acknowledgements

We thank F. Marchand and M. Convertino for interesting discussions, and S. Muff for comments on the manuscript. The simulations were performed on the Matterhorn cluster of the University of Zurich, and we gratefully acknowledge the support of C. Bolliger and A. Godknecht. This work was supported by a Swiss National Science Foundation grant and the National Competence Center for Research (NCCR) in Neural Plasticity and Repair.

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2007.09.090

References

1. Dobson, C. M. (2003). Protein folding and misfolding. *Nature*, **426**, 884–890.
2. Lansbury, P. T. & Lashuel, H. A. (2006). A century-old debate on protein aggregation and neurodegeneration enters the clinic. *Nature*, **443**, 774–779.
3. Haass, C. & Selkoe, D. J. (2007). Soluble protein oligomers in neurodegeneration: lessons from the Alzheimer's amyloid β-peptide. *Nature Rev. Mol. Cell Biol.* **8**, 101–112.
4. Caughey, B. & Lansbury, P. T. (2003). Protofibrils, pores, fibrils, and neurodegeneration: separating the responsible protein aggregates from the innocent bystanders. *Annu. Rev. Neurosci.* **26**, 267–298.
5. Nilsberth, C., Westlind-Danielsson, A., Eckman, C. B., Condron, M. M., Axelman, K., Forsell, C. *et al.* (2001). The 'Arctic' APP mutation (E693G) causes Alzheimer's disease by enhanced Abeta protofibril formation. *Nature Neurosci.* **4**, 887–893.
6. Cheng, I. H., Palop, J. J., Esposito, L. A., Bien-Ly, N., Yan, F. & Mucke, L. (2004). Aggressive amyloidosis in mice expressing human amyloid peptides with the Arctic mutation. *Nature Med.* **10**, 1190–1192.
7. Conway, K. A., Lee, S. J., Rochet, J. C., Ding, T. T., Williamson, R. E. & Lansbury, P. T. (2000). Acceleration of oligomerization, not fibrillization, is a shared property of both alpha-synuclein mutations linked to early-onset Parkinson's disease: implications for pathogenesis and therapy. *Proc. Natl Acad. Sci. USA*, **97**, 571–576.
8. Lomakin, A., Chung, D. S., Benedek, G. B., Kirschner, D. A. & Teplow, D. B. (1996). On the nucleation and growth of amyloid β-protein fibrils: detection of nuclei and quantitation of rate constants. *Proc. Natl Acad. Sci. USA*, **93**, 1125–1129.
9. O'Nuallain, B., Shivaprasad, S., Kheterpal, I. & Wetzel, R. (2005). Thermodynamics of Aβ(1-40) amyloid fibril elongation. *Biochemistry*, **44**, 12709–12718.
10. Kelly, J. W. (1998). The alternative conformations of amyloidogenic proteins and their multi-step assembly pathways. *Curr. Opin. Struct. Biol.* **8**, 101–106.
11. Gosal, W. S., Morten, I. J., Hewitt, E. W., Smith, D. A., Thomson, N. H. & Radford, S. E. (2005). Competing pathways determine fibril morphology in the self-assembly of β2-microglobulin into amyloid. *J. Mol. Biol.* **351**, 850–864.
12. Jahn, T. R., Parker, M. J., Homans, S. W. & Radford, S. E. (2006). Amyloid formation under physiological conditions proceeds *via* a native-like folding intermediate. *Nature Struct. Mol. Biol.* **13**, 195–201.
13. Lomakin, A., Teplow, D. B., Kirschner, D. A. & Benedek, G. (1997). Kinetic theory of fibrillogenesis of amyloid β-protein. *Proc. Natl Acad. Sci. USA*, **94**, 7942–7947.
14. Massi, F. & Straub, J. E. (2001). Energy landscape theory for Alzheimer's amyloid β-peptide fibril elongation. *Proteins: Struct. Funct. Bioinformatics*, **42**, 217–229.
15. Modler, A. J., Gast, K., Lutsch, G. & Damaschun, G. (2003). Assembly of amyloid protofibrils *via* critical oligomers–a novel pathway of amyloid formation. *J. Mol. Biol.* **325**, 135–148.
16. Hall, D., Hirota, N. & Dobson, C. M. (2005). A toy model for predicting the rate of amyloid formation from unfolded protein. *J. Mol. Biol.* **351**, 195–205.
17. Carulla, N., Caddy, G. L., Hall, D. R., Zurdo, J., Gairi, M., Feliz, M. *et al.* (2005). Molecular recycling within amyloid fibrils. *Nature*, **436**, 554–558.
18. Broglia, R. A., Tiana, G., Pasquali, S., Roman, H. E. & Vigezzi, E. (1998). Folding and aggregation of designed proteins. *Proc. Natl Acad. Sci. USA*, **95**, 12930–12933.
19. Gupta, P., Hall, C. K. & Voegler, A. C. (1998). Effect of denaturant and protein concentrations upon protein refolding and aggregation: a simple lattice model. *Protein Sci.* **7**, 2642–2652.
20. Harrison, P. M., Chan, H. S., Prusiner, S. B. & Cohen, F. E. (1999). Thermodynamics of model prions and its implications for the problem of prion protein folding. *J. Mol. Biol.* **286**, 593–606.
21. Dima, R. I. & Thirumalai, D. (2002). Exploring protein aggregation and self-propagation using lattice models: phase diagram and kinetics. *Protein Sci.* **11**, 1036–1049.
22. Urbanc, B., Cruz, L., Yun, S., Buldyrev, S. V., Bitan, G., Teplow, D. B. & Stanley, H. E. (2004). In silico study of amyloid β-protein folding and oligomerization. *Proc. Natl Acad. Sci. USA*, **101**, 17345–17350.
23. Nguyen, H. D. & Hall, C. K. (2004). Molecular

dynamics simulations of spontaneous fibril formation by random-coil peptides. *Proc. Natl Acad. Sci. USA*, **101**, 16180–16185.
24. Jang, H., Hall, C. K. & Zhou, Y. (2004). Assembly and kinetic folding pathways of a tetrameric beta-sheet complex: molecular dynamics simulations on simplified off-lattice protein models. *Biophys. J.* **86**, 31–49.
25. Khare, S. D., Ding, F., Gwanmesia, K. N. & Dokholyan, N. V. (2005). Molecular origin of polyglutamine aggregation in neurodegenerative diseases. *PLoS Comput. Biol.* **1**, 230–235.
26. Chen, Y. & Dokholyan, N. V. (2005). A single disulfide bond differentiates aggregation pathways of beta2-microglobulin. *J. Mol. Biol.* **354**, 473–482.
27. Malolepsza, E., Boniecki, M., Kolinski, A. & Piela, L. (2005). Theoretical model of prion propagation: a misfolded protein induces misfolding. *Proc. Natl Acad. Sci. USA*, **102**, 7835–7840.
28. Bellesia, G. & Shea, J.-E. (2007). Self-assembly of beta-sheet forming peptides into chiral fibrillar clusters. *J. Chem. Phys.* **126**, 245104.
29. Plakoutsi, G., Bemporad, F., Calamai, M., Taddei, N., Dobson, C. M. & Chiti, F. (2005). Evidence for a mechanism of amyloid formation involving molecular reorganisation within native-like precursor aggregates. *J. Mol. Biol.* **351**, 910–922.
30. Ma, B. & Nussinov, R. (2002). Stabilities and conformations of Alzheimer's β-amyloid peptide oligomers ($A\beta_{16-22}$, $A\beta_{16-35}$, and $A\beta_{10-35}$): sequence effects. *Proc. Natl Acad. Sci. USA*, **99**, 14126–14131.
31. Gsponer, J., Haberthür, U. & Caflisch, A. (2003). The role of side-chain interactions in the early steps of aggregation: molecular dynamics simulations of an amyloid-forming peptide from the yeast prion Sup35. *Proc. Natl Acad. Sci. USA*, **100**, 5154–5159.
32. Klimov, D. & Thirumalai, D. (2003). Dissecting the assembly of $A\beta_{16-22}$ amyloid peptides into antiparallel β sheets. *Structure*, **11**, 295–307.
33. Wei, G., Mousseau, N. & Derreumaux, P. (2004). Sampling the self-assembly pathways of KFFE hexamers. *Biophys. J.* **87**, 3648–3656.
34. Hwang, W., Zhang, S., Kamm, R. D. & Karplus, M. (2004). Kinetic control of dimer structure formation in amyloid fibrillogenesis. *Proc. Natl Acad. Sci. USA*, **101**, 12916–12921.
35. Buchete, N.-V., Tycko, R. & Hummer, G. (2005). Molecular dynamics simulations of Alzheimer's beta-amyloid protofilaments. *J. Mol. Biol.* **353**, 804–821.
36. Lopez de la Paz, M., de Mori, G. M. S., Serrano, L. & Colombo, G. (2005). Sequence dependence of amyloid fibril formation: insights from molecular dynamics simulations. *J. Mol. Biol.* **349**, 583–596.
37. Pellarin, R. & Caflisch, A. (2006). Interpreting the aggregation kinetics of amyloid peptides. *J. Mol. Biol.* **360**, 882–892.
38. Kodali, R. & Wetzel, R. (2007). Polymorphism in the intermediates and products of amyloid assembly. *Curr. Opin. Struct. Biol.* **17**, 48–57.
39. Murphy, R. M. (2007). Kinetics of amyloid formation and membrane interaction with amyloidogenic proteins. *Biochim. Biophys. Acta*, **1768**, 1923–1934.
40. Rao, F. & Caflisch, A. (2004). The protein folding network. *J. Mol. Biol.* **342**, 299–306.
41. Caflisch, A. (2006). Network and graph analyses of folding free energy surfaces. *Curr. Opin. Struct. Biol.* **16**, 71–78.
42. Esler, W. P., Stimson, E. R., Jennings, J. M., Vinters, H. V., Ghilardi, J. R., Lee, J. P. *et al.* (2000). Alzheimer's disease amyloid propagation by a template-dependent dock-lock mechanism. *Biochemistry*, **39**, 6288–6295.
43. Gobbi, M., Colombo, L., Morbin, M., Mazzoleni, G., Accardo, E., Vanoni, M. *et al.* (2006). Gerstmann-Sträussler-Scheinker disease amyloid protein polymerizes according to the "dock-and-lock" model. *J. Biol. Chem.* **281**, 843–849.
44. Nguyen, P. H., Li, M. S., Stock, G., Straub, J. E. & Thirumalai, D. (2007). Monomer adds to preformed structured oligomers of abeta-peptides by a two-stage dock-lock mechanism. *Proc. Natl Acad. Sci. USA*, **104**, 111–116.
45. Jansen, R., Dzwolak, W. & Winter, R. (2005). Amyloidogenic self-assembly of insulin aggregates probed by high resolution atomic force microscopy. *Biophys. J.* **88**, 1344–1353.
46. Walsh, D., Hartley, D. M., Kusumoto, Y., Fezoui, Y., Condron, M., Lomakin, A. *et al.* (1999). Amyloid β-protein fibrillogenesis. structure and biological activity of protofibrillar intermediates. *J. Biol. Chem.* **274**, 25945–25952.
47. Fersht, A. (1999). *Structure and Mechanism in Protein Science*. W.H. Freeman and Company, New York, NY.
48. Tartaglia, G. G., Cavalli, A., Pellarin, R. & Caflisch, A. (2004). The role of aromaticity, exposed surface, and dipole moment in determining protein aggregation rates. *Protein Sci.* **13**, 1939–1941.

7 Computer-aided stabilization of the hydrophobic core of a consensus designed repeat protein.

Globular soluble proteins usually fold into a compact structure where hydrophobic side-chains are partitioned in the interior and polar residue exposed on the surface of the molecule. The force that drives a disordered polypeptide to achieve a collapsed structure is called the hydrophobic effect [76, 77]. Crystal structures of proteins display tightly packed side-chains in the core, an arrangement that provides a favorable van der Waals energy. However, even a tight packing is not robust upon mutagenesis: a single mutation in the inside of the protein can be highly destabilizing and might force the protein to adopt many alternative conformations with similar energies. Under these conditions, if the secondary structure of the mutated protein is preserved, then it is said that the protein has molten globule-like features. A molten globule [78, 79] is a particular state that characterize proteins under mildly denaturing conditions. They are characterized by a compact state and a "fluid" hydrophobic core, lacking specific native interactions. Presence of secondary structure, poor signal dispersion in NMR spectra and high affinity for hydrophobic dyes are typical molten globule features, indicating an intermediate behavior between folded and unstructured proteins [80]. The unfolding reaction scheme expressed by equation (1) can be generalized in presence of a on-pathway intermediate:

$$F \rightleftharpoons I \rightleftharpoons U \tag{9}$$

Transition from a denaturated state U to a molten globule I is thought to be a continuous transition, lacking of cooperativity. Conformational stability of proteins, measured by thermal denaturation, can also be triggered by mutations of aliphatic amino acids that are in the hydrophobic core [81]. In this work the hydrophobic optimization of a consensus designed armadillo repeat protein (see section 7.1) that shows molten-globule characteristics, is approached using a structure-based engineering, with the aim of producing a modular protein with specific binding properties [82]. Consensus design is a statistical analysis of sequence alignments of families of homologous proteins, which is used to improve stability of proteins with respect to the natural sequences [83]. One dif-

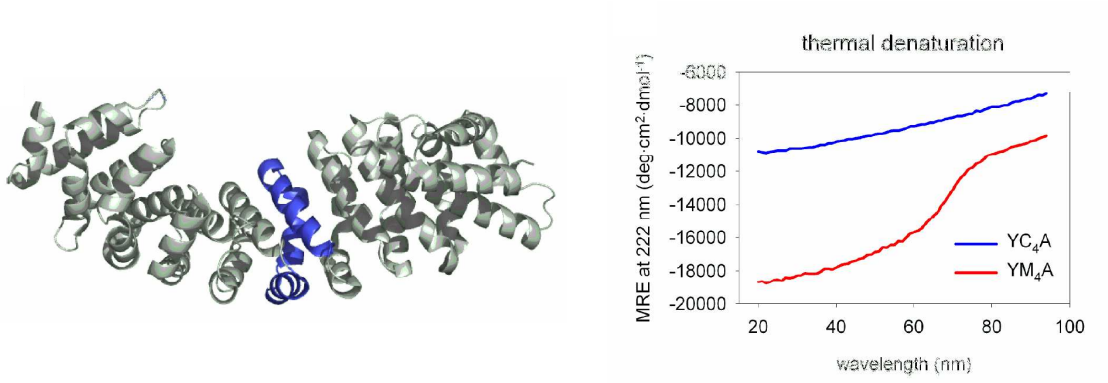


Figure 5: Left: employed scaffold for the computational design. The blue ribbon is the unit repeated module. Right: thermal stability of wild type (blue) vs selected mutant (red). CD ellipticity at 222 nm is measured at different temperatures.

ficulty in the *de novo* protein design is achieving the side chain packing needed to create a stable native state. It has been reported that consensus design might seldom fail in the engineering of hydrophobic cores [84] leading to structures that have loose activity, or are prone to aggregate.

A number of computational techniques have been employed for protein engineering [85, 86], especially in the design of the hydrophobic core [87, 88]. The final task of computer-aided protein design is to search, within a selected conformational space of the protein, the global energy minimum. The efficiency of this search is limited by the number of high energy minima, which increases exponentially with the protein size. In the case presented in section 7.1, due to the size of the system (about 600 residues), a number of approximations were necessary. A restricted pool of 432 hydrophobic core mutants were selected, using the most frequent amino acid substitutions of the consensus rank. The calculations have been executed on three scaffolds adapted from X-ray structures (see figure 5), in vacuum. For each mutant a random sampling of the inner side-chains rotamers was performed, and the force field energy was used to evaluate and rank every single mutant. Therefore the 20 most promising mutants were employed for further experimental investigation. One of these mutations considerably enhanced the wild type stability, displaying a greater solubility, compactness, secondary structure, folding cooperativity (see figure 5) and a broad NMR spectra. The overall performance

of the computational design is very high, in fact all of the 20 highest ranked mutants show equal or better characteristics than the wild type.

7.1 Designed armadillo repeat proteins as general peptide-binding scaffolds: consensus design and computational optimization of the hydrophobic core [J. Mol. Biol. 2008, *376*, 1282]

doi:10.1016/j.jmb.2007.12.014 J. Mol. Biol. (2008) 376, 1282–1304

JMB

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Designed Armadillo Repeat Proteins as General Peptide-Binding Scaffolds: Consensus Design and Computational Optimization of the Hydrophobic Core

Fabio Parmeggiani[1], Riccardo Pellarin[1], Anders Peter Larsen[1], Gautham Varadamsetty[1], Michael T. Stumpp[1], Oliver Zerbe[2], Amedeo Caflisch[1] and Andreas Plückthun[1]*

[1]*Department of Biochemistry, University of Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland*

[2]*Department of Organic Chemistry, University of Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland*

Received 3 July 2007; received in revised form 13 November 2007; accepted 5 December 2007
Available online 14 December 2007

Edited by F. E. Cohen

Armadillo repeat proteins are abundant eukaryotic proteins involved in several cellular processes, including signaling, transport, and cytoskeletal regulation. They are characterized by an armadillo domain, composed of tandem armadillo repeats of approximately 42 amino acids, which mediates interactions with peptides or parts of proteins in extended conformation. The conserved binding mode of the peptide in extended form, observed for different targets, makes armadillo repeat proteins attractive candidates for the generation of modular peptide-binding scaffolds. Taking advantage of the large number of repeat sequences available, a consensus-based approach combined with a force field-based optimization of the hydrophobic core was used to derive soluble, highly expressed, stable, monomeric designed proteins with improved characteristics compared to natural armadillo proteins. These sequences constitute the starting point for the generation of designed armadillo repeat protein libraries for the selection of peptide binders, exploiting their modular structure and their conserved binding mode.

Keywords: consensus design; armadillo repeat; hydrophobic core; molten globule; molecular dynamics and minimization

Introduction

In recent years, as an alternative to raising monoclonal antibodies by immunization, recombinant antibodies[1] and an increasing number of other protein scaffolds[2] have been investigated as novel binding molecules. However, neither antibodies themselves nor any of these alternative protein scaffolds were specifically designed to bind peptides. Target-specific binding molecules are, in general, obtained from large protein libraries by *in vitro* selection or, in the case of monoclonal antibodies, through traditional immunization procedures. Both approaches require that, for each target, each new binding molecule is individually generated and characterized for specificity and

*Corresponding author. E-mail address: plueckthun@bioc.uzh.ch.

Present addresses: A.P. Larsen, Department of Biomedical Sciences, University of Copenhagen, Blegdamsvej 3, DK-2200 Copenhagen, Denmark; M.T. Stumpp, Molecular Partners AG, Grabenstrasse 11a, CH-8952 Schlieren, Switzerland.

Abbreviations used: αArm, Armadillo domain of human importin-α1 (residues 83–505); βArm, Armadillo domain of mouse β-catenin (residues 150–665); ANS, 1-anilino-naphthalene-8-sulfonate; C-type, overall consensus repeat; CD, circular dichroism; HA, hemagglutinin tag; HSQC, heteronuclear single quantum coherence; I-type, importin-derived consensus armadillo repeat; IMAC, immobilized metal-ion affinity chromatography; M-type, mutated armadillo repeat obtained by computational approach; MALS, multiangle light scattering; MRE, mean residue ellipticity; NLS, nuclear localization sequence; NOE, nuclear Overhauser enhancement; pD, phage lambda protein D; PDB, Protein Data Bank; SDS-PAGE, sodium dodecyl sulfate–polyacrylamide gel electrophoresis; SEC, size-exclusion chromatography; T-type, catenin/plakoglobin-derived consensus armadillo repeat.

0022-2836/$ - see front matter

cross-reactivity, making the generation of binders against a large number of peptide targets (e.g., representing a full proteome) an almost prohibitive task.

The aim of the present study was to develop a scaffold for the generation of peptide-specific binding proteins. In more detail, we wanted to develop proteins that were stable under various conditions and with the intrinsic ability to bind peptides in a conserved fashion. To recognize peptides in a sequence-selective manner the specificity of binding should ideally be conferred through specific interactions with the peptide side chains.

Natural peptide-binding scaffolds can be grouped in different classes. Antibodies are known to be able to bind peptides and have been well characterized.[3–6] Although peptide-binding antibodies have certain structural features in common, the mode of binding is not conserved. Thus, the information acquired through studies of antibody–peptide complexes cannot easily be applied to the general design of peptide-binding antibodies or extended to other proteins.

Small adaptor domains (e.g., SH2, SH3, and PDZ)[7] show specific binding to their targets, usually in a conserved binding mode within one family, but their affinity is generally low. The recognition sequence is very short and biased toward certain amino acid types, posttranslational modifications, or free N- or C-termini. While several such domains could be linked together by flexible peptides to recognize longer peptide sequences, a coverage of any arbitrary sequence would still be very difficult since these small domains might not be adaptable to the recognition of any arbitrary sequence. Furthermore, the entropy loss upon binding of such flexibly linked constructs would not necessarily lead to high affinities.

The major histocompatibility complex proteins (MHC I and MHC II)[8] possess a higher intrinsic variability and the ability to recognize a broad range of peptides, but the difficulties in their handling reduce their attractiveness as a scaffold candidate.

Repeat proteins, in particular tetratricopeptide repeats (TPRs),[9] armadillo,[10] and WD40[11] proteins, have been shown to possess an intrinsic ability to bind peptides, taking advantage of their repetitive structure. Thus, for our purpose, a scaffold based on repeat proteins seemed to constitute a promising candidate. For reasons outlined below, we chose the armadillo repeat protein family as the basis for our scaffold candidate.

Armadillo repeat proteins[12,13] are abundant in eukaryotes, where they are involved in a broad range of biological processes (e.g., transcription regulation,[14] cell adhesion,[15] tumor suppressor activity,[16] and nucleocytoplasmic transport[17]). These proteins are characterized by tandem repeats of approximately 42 amino acids that were first discovered in the product of the *Drosophila melanogaster* segmentation polarity gene Armadillo, which is homologous to mammalian β-catenin.[18,19] Armadillo repeat proteins participate in protein–protein interactions, and the armadillo domain is usually involved in the recognition process. The domain forms a right-handed superhelix[20,21] (Fig. 1a), as shown by the crystal structures of β-catenin[22] and importin-α.[23] Every repeat is composed of three α-helices, named H1, H2, and H3 (Fig. 1b), and several repeats stack to form the compact domain. Specialized repeats are present at the N- and C-termini of the protein, protecting the hydrophobic core from solvent exposure (Fig. 1a).

Armadillo repeat proteins are able to bind different types of peptides, yet relying on a conserved binding mode of the peptide backbone. Reported dissociation constant (K_d) values as low as 10–20 nM[24]

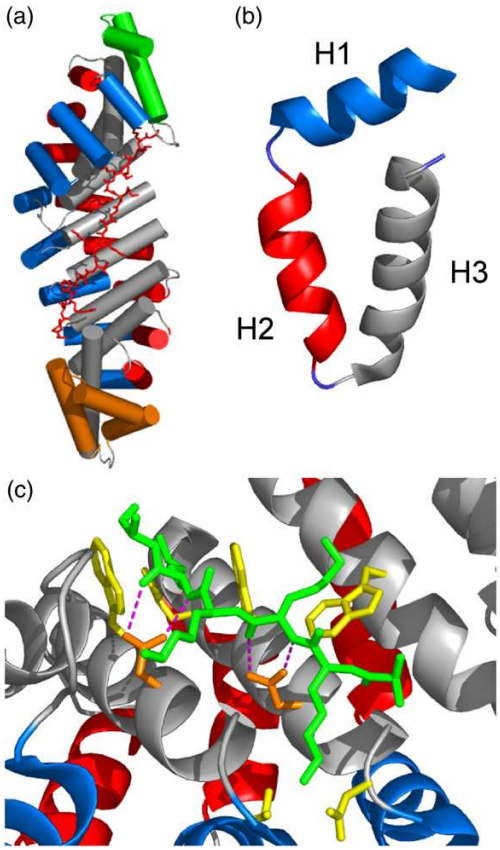


Fig. 1. (a) Structure of *S. cerevisiae* importin-α in complex with nucleoplasmin NLS (PDB ID 1EE5), showing the right-handed superhelical structure typical for armadillo repeat proteins. The cylinders represent the α-helices. The N-terminal repeat is indicated in green, and the C-terminal repeat is shown in orange. The bound peptide is depicted in red in a stick representation. (b) Detail of repeat 6 from 1EE5. The α-helices are represented as ribbons. (c) Detail of the peptide-binding mode. The conserved asparagine residues (in orange) contact *via* hydrogen bonds (purple) the backbone of the peptide, depicted in green. The residues that are responsible for the interactions with the side chains of the target peptide are shown in yellow. In all panels, helix 1 (H1) is indicated in blue, helix 2 (H2) in red, and helix 3 (H3) in gray.

indicate that high affinities can be achieved. Crystal structures of armadillo repeat proteins in complex with bound peptides have revealed that most peptide targets are bound in an extended conformation along the surface, inside the groove formed by the H3 helices. The superhelical armadillo domain winds around the peptide, oriented in the opposite N- to C-terminal direction (Fig. 1a), thus forming a double-helical complex, topologically similar to the DNA double strand. An asparagine residue, conserved in almost every repeat at the C-terminal part of H3, makes hydrogen bonds to the main chain of the target peptide, thereby keeping it in an extended conformation. Additional interactions to the target side chains are provided by neighboring residues, mostly in H3 (Fig. 1c). In a first approximation, each dipeptide unit of the target peptide is specifically recognized by one repeat in the armadillo domain (Fig. 2a).

In theory, the possibility of developing individual repeats that specifically bind a two-amino-acid sequence unit is very attractive. Given that the individual repeats are based on the same optimized scaffold and, thus, compatible with each other, any given number of repeats can be directly stacked to extend the recognition to much longer peptide sequences. In contrast to flexibly linked small adaptor domains mentioned above, armadillo repeats directly stack on each other in a rather rigid manner, allowing binding to uninterrupted longer peptides. This would exploit the specificity of the individual repeats to provide a peptide-binding designed armadillo protein with high and predetermined specificity, governed by the individual repeats. Such an approach (Fig. 2b), using armadillo proteins assembled from previously selected "building blocks," could effectively bypass the current *in vitro* selection procedures for individual peptides. However, this requires such individual peptide-specific repeats first to be developed, using a library-based approach.

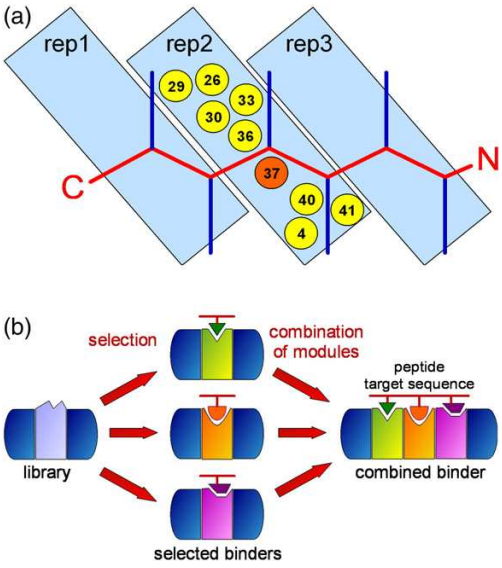


Fig. 2. Binding of target peptides. (a) Schematic drawing of an armadillo repeat protein binding to an extended peptide. The target peptide is bound in an antiparallel orientation to the protein. N and C indicate N- and C-termini of the peptide, which is depicted in red, with the amino acid side chains shown in blue. The residues of armadillo repeats involved in binding occupy specific positions within the single repeat sequences, mostly on helix 3. The position indicated in orange (a conserved Asn) is responsible for the binding of the peptide main chain; the positions in yellow are involved in recognition of the peptide side chains. (b) Designed armadillo repeat proteins potentially allow the selection of single repeats that specifically recognize short sequences. The selected peptide-specific repeats can be then combined to recognize longer peptides without performing additional selections.

In the present study, we have, as a first step, designed armadillo repeat modules based on consensus sequences. Proteins containing different types of modules have been assembled and characterized, initially only leading to stable dimeric proteins or monomeric molten-globule-like proteins. We subsequently used a combination of molecular dynamics and minimization to improve the hydrophobic core packing and convert the consensus-designed armadillo repeat protein with molten-globule-like properties to a monomeric, stable folded protein. Finally, the protein characteristics were evaluated for exploring the possibility of generating a modular peptide-binding scaffold. We succeeded in developing a stable, monomeric consensus protein that can be used now in the generation of peptide-specific individual armadillo repeat proteins.

Results

Armadillo repeat protein design

A consensus design strategy[25] has been applied in order to generate armadillo repeat proteins with high expression levels of soluble protein in *Escherichia coli*, monomeric state, high thermodynamic stability, and absence of cysteines for convenient expression and handling.

This design procedure was aimed at the generation of self-compatible repeat modules; therefore, consensus sequences were derived from multiple alignments of single armadillo repeats from the Swiss-Prot database.[26] A consensus design strategy has been successfully applied previously to other designed repeat proteins,[27–30] and it is based on consensus design of internal repeats (or internal modules). Special terminal capping repeats (terminal modules) have been generated to protect the hydrophobic core from solvent exposure. The crucial role of capping repeats has been previously shown in studies with designed ankyrin repeat proteins.[31]

The numbering used here to define the positions within the repeats was based on the family align-

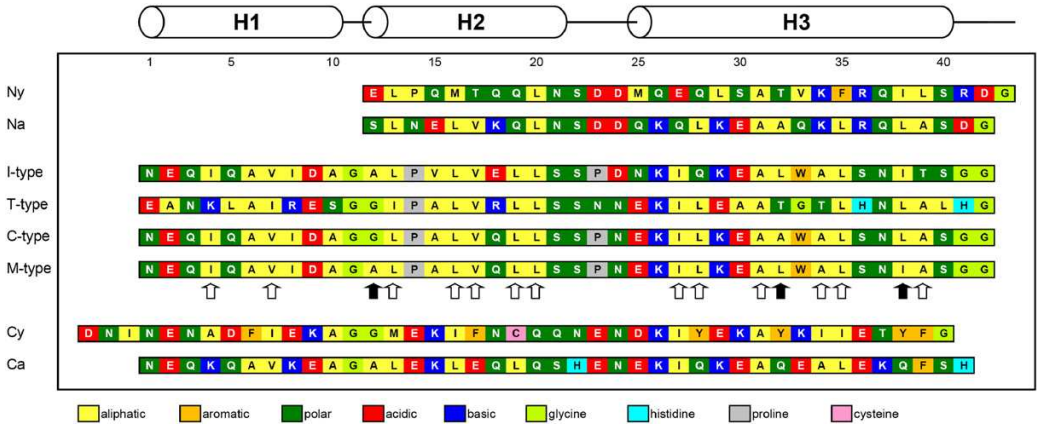


Fig. 3. Sequences of the designed internal modules and capping repeats; the cylinders indicate the helices, and the numbers denote the positions within the single repeats according to the convention introduced. *Ny* is an N-terminal capping repeat derived from importin-α from the *y*east *S. cerevisiae*; *Na* is an *a*rtificial N-terminal capping repeat. *I-type* is the internal module based on sequences from the *i*mportin-α subfamily, *T-type* is the internal module based on sequences from the β-cat*e*nin/plakoglobin subfamily, *C-type* is the overall *c*onsensus based on both subfamilies. *M-type* is the mutant sequence obtained through the computational approach described here. *Cy* is a C-terminal capping repeat derived from importin-α from the *y*east *S. cerevisiae*; *Ca* is an *a*rtificial C-terminal capping repeat. The amino acid color code is indicated below the sequences. The arrows indicate the positions considered in the computational approach. The filled arrows show the positions that differ between the C-type and the finally chosen M-type module.

ment proposed by Andrade *et al.*[32] Position −2 in that work corresponds to position 1 in the numbering used in the present study, where the putative helices H1, H2, and H3 encompass residues 1–10, 12–21, and 25–40, respectively.

Consensus design of internal modules

The initial sequence profile was generated using the family alignment from SMART†[33,34] (data from January 2004) as starting point (the consensus sequence is shown in Supplementary Fig. S1a). We largely followed the steps previously outlined.[27,30] We first removed all sequences lacking annotation in the Swiss-Prot database, especially hypothetical proteins or sequences for which no protein data were available with the exception of indirect evidence by sequence homology. The final set of 319 sequences led to a profile of 40 residues, covering the repeat sequence from H1 to H3 but excluding the loop between H3 and the next repeat. This sequence profile was used for a further search against the Swiss-Prot database. The repeats thus found belong to proteins that fall into different subfamilies of armadillo repeat proteins, as indicated by Andrade *et al.*[32]

Nevertheless, the sequences from different subfamilies might not be compatible. Taking this possibility into account, three final consensus sequences were constructed: one derived from β-catenin/plakoglobin (110 sequences), one from importin-α (133 sequences), and one from the combination of both (243 sequences). No normalization was applied during the calculation of the combined consensus sequence, which would compensate for a slight overrepresentation of importin-α over β-catenin/plakoglobin sequences in our selected set (133 over 110). The automatic alignments, performed with ClustalW[35] (Supplementary Fig. S1b), were manually refined including the loop connecting adjacent repeats (Supplementary Fig. S1c).

Structural information was taken into account to replace the cysteines present in the consensus sequences and reduce possible steric clashes. A more detailed description of additional sequence features and the rationale for amino acid exchanges are provided in the Supplementary Materials. Requirements for the cloning strategy were also considered at this stage, leading to the final module sequences type I (derived from *i*mportin-α subfamily), type T (derived from β-cat*e*nin/plakoglobin subfamily), and type C (*c*ombined consensus between these two subfamilies) (Fig. 3). Positions 7, 16, 17, 19, 20, 31, 34, 35, and 38 are well conserved in all the sequences and are part of the hydrophobic core of the armadillo proteins.

The positions potentially involved in binding of peptides (4, 26, 29, 30, 33, 36, 37, 40, and 41) have been defined based on the analysis of structures of complexes (summarized by Lange *et al.*[36] and Xu and Kimelman[37]) and data from mutation experiments.[38–40] The conserved Asn, responsible for binding to the main chain of the target peptide and at least in part for keeping it in an extended conformation, is located at position 37. Position 4 is part of both the hydrophobic core and the peptide binding site, and thus, the types of residues allowed at this position in a potential library would probably be restricted.

† http://smart.embl.de

Design of capping repeats

N- and C-terminal capping repeats, found in natural armadillo domains, protect the hydrophobic core, as they present a hydrophobic surface to the internal repeat side but a hydrophilic surface to the solvent. Capping sequences have also been considered in the previous design of other repeat proteins.[27,29–31]

The boundaries of armadillo domains have been estimated by limited proteolysis.[22,23] However, they are not clearly defined, partly due to the weak similarity of the terminal repeats to the internal ones. In addition, not all the residues are visible in the crystal structures of importin-α and β-catenin. It is likely that only the visible residues contribute to the armadillo domain, and the additional parts are unstructured and do not strictly belong to the domain. We have defined the N-terminal capping repeat as starting from position 12 (the beginning of H2). In contrast, the C-terminal capping repeat is completely resolved in the x-ray structures, and we defined it to comprise position 1 to position 41, thus including H1 to H3.

The capping repeats have been designed by using two different approaches. In the first, natural capping repeats were adapted to our designed internal repeats. Structural information to ensure compatibility between the capping repeats and the designed internal repeat is a fundamental prerequisite. The importin-α from *Saccharomyces cerevisiae* was considered to be the best candidate for a general capping repeat donor: all our designed modules present a flat surface that can interact with the inner surface of yeast importin-α capping repeats, as judged from molecular models. The *y*east importin-α-derived N-terminal and C-terminal capping repeats were named N*y* and C*y*, respectively.

The N-terminal capping repeat covers the residues from Glu88 to His119 of yeast importin-α. However, the two residues Glu118–His119 were replaced by Asp–Gly (Fig. 3, positions 42 and 43 of Ny) to adapt the terminal loop to the designed modules: glycine is used for assembly of the modules (as its codon overlaps a restriction site) and aspartate keeps a negative charge, which is frequently present at this position in natural proteins, reducing at the same time the helical propensity in the turn region.

The C-terminal capping repeat covers the region from Asn471 to Gly510 in yeast importin-α. However, the loop connecting the last internal repeat with the C-terminal repeat contains additional residues in yeast importin-α, compared to other natural importins. A modified version of this C-terminal capping repeat has thus been generated by introducing three residues (Asp-Asn-Ile) before H1 (Fig. 3, first three residues of Cy). Asn and Ile are naturally present at these positions; Asp has been included to keep a negatively charged loop as observed in several natural sequences while reducing the helical propensity.

In the second approach, two completely *a*rtificial N- and C-terminal capping repeats were designed (named N*a* and C*a*, respectively, and shown in Fig. 3), starting from the type C consensus and substituting the exposed hydrophobic residues with hydrophilic ones. Positions 12, 19, 27, and 34 of the N-terminal capping repeat are occupied by hydrophobic residues in the consensus sequence and were replaced by hydrophilic residues based on structures and frequently occurring residues obtained from alignment of N-terminal capping sequences. In a similar way, positions 8, 13, 17, 20, 28, 32, 35, 38, and 39 of the C-terminal capping repeat were replaced by hydrophilic residues based on structures and frequently occurring residues obtained from alignment of C-terminal capping sequences. A detailed description of the residues introduced in the designed capping repeats is provided in the Supplementary Materials. A second version of the Cy capping repeat was also designed without the three initial residues and with Ala replacing Cys at position 19; no change was observed, compared to Cy, in the level of expression and in the amount of soluble protein, and it will thus not be discussed further.

Assembly, cloning, and expression of designed armadillo repeat proteins

The amino acid sequences of all modules were back-translated to DNA sequences, optimizing the codon usage for expression in *E. coli*. Each module was synthesized, starting from overlapping oligonucleotides (Supplementary Table S1).

The modules were assembled stepwise using type IIS restriction enzymes (Supplementary Fig. S2), following the approach reported by Binz *et al.*[27] The final proteins were named according to the modules that they contain: the name indicates, in order, the type of N-terminal repeat (A for *A*rtificial or N*a*, Y for *Y*east derived or N*y*), the type of internal repeats (type I, type T, or type C) with the number of modules used as a subscript, and the type of C-terminal repeat (A for *A*rtificial or C*a*, Y for *Y*east derived or C*y*): for example, YI_4A contains a *Y*east-derived N-terminal repeat (N*y*), four internal repeats based on *I*mportin consensus (type *I*), and an *A*rtificial C-terminal repeat (C*a*).

Thus, Na or Ny as N-capping modules were combined with T-, I-, or C-type internal modules and Ca or Cy C-terminal modules, leading to 12 possible combinations. The proteins contain only one type of internal module to avoid incompatible surfaces at the interface between repeats. The influence of capping and internal repeats was evaluated by analyzing the expression properties of all the constructs, containing two or four internal repeats. The proteins were expressed in *E. coli* XL1-blue using a pQE30-based expression plasmid, providing an N-terminal $MRGSH_6$ tag for purification. The insert was constructed with a double stop codon (Supplementary Fig. S3). As an example, the DNA and protein sequences of YC_2A are provided in Supplementary Fig. S3.

The highest level of soluble protein expression was obtained when the internal modules were combined with Ny and Ca (Fig. 4a). The Na cap leads to almost undetectable expression in Coomassie-stained polyacrylamide gels, and the presence of Cy resulted in a substantial portion of the protein found in the insoluble fraction after cell lysis. The observed effects of terminal capping repeats were independent of the type and the number of internal modules. However, increasing the number of internal modules enhanced the amount of soluble protein and the absolute amount of protein produced. Remarkably, type T proteins are characterized by a lower apparent mobility in sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS-PAGE), compared to type I and type C proteins (Fig. 4b).

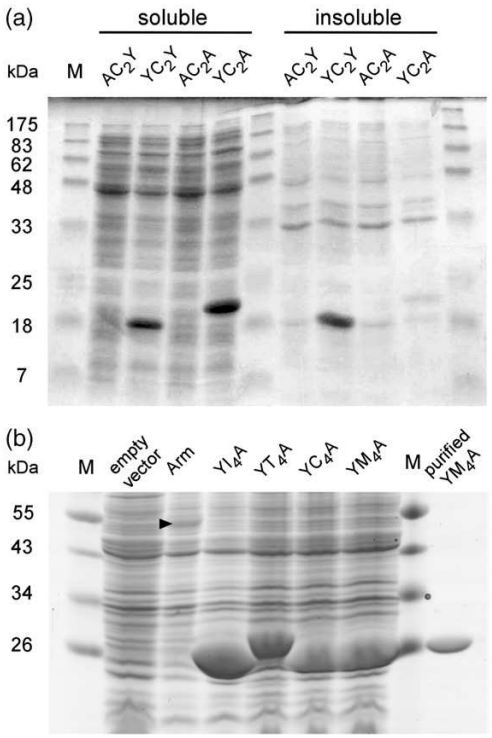


Fig. 4. (a) Influence of capping repeats on expression. Soluble and insoluble fractions of *E. coli* cell extracts are shown in a Coomassie-stained SDS polyacrylamide gel. The proteins contain two internal C-type modules with different combinations of capping repeats. (b) Whole-cell extracts of consensus proteins. The constructs contain Ny and Ca as capping repeats. Cells transformed with the empty vector or with the vector containing the armadillo domain of mouse β-catenin (Arm) were used as control. The proteins can be easily purified in a single step by IMAC, as shown for YM_4A. The expected size is 18 and 27 kDa for proteins containing two or four internal modules, respectively, and 56 kDa for Arm. The triangle indicates the band corresponding to the armadillo domain of β-catenin, which is expressed at much lower yield than the designed proteins. The molecular mass of the marker (M) is indicated in kilodaltons on the left.

Protein purification and characterization: Comparison with natural armadillo domains

Proteins containing the combination of Ny, Ca, and two, four, or eight internal repeats have been chosen for biophysical characterization and evaluation of the properties of type I, type T, and type C modules. The results are summarized in Table 1.

The purification by immobilized metal-ion affinity chromatography (IMAC) in a single step provided up to 100 mg of pure protein from 1 l of bacterial culture (Fig. 4b). No sign of precipitation or degradation was detected by spectrophotometry and SDS-PAGE in protein solutions stored for up to 1 month at 4 °C in the IMAC elution buffer.

The natural human importin-α1 (Swiss-Prot P52294) and the mouse β-catenin (Swiss-Prot Q02248) were also expressed using the same pQE30-based plasmid. The importin contains 10 armadillo repeats and the catenin 12, including the capping repeats in the count in both cases. Human importin-α1 gave the highest yield among the importin-α family members tested (data not shown) after a two-step coupled IMAC–ion exchange purification and, together with mouse β-catenin, was used for the comparison with the designed armadillo repeat proteins.

Importin-α1 and β-catenin, despite their elongated shape, elute at the volume expected from their molecular weight in gel filtration on a Superdex 200 column, and the monomeric state was confirmed for both proteins by multiangle light scattering (MALS) measurements (Table 1).

On the other hand, the designed proteins show elution volumes corresponding to higher-than-expected apparent molecular masses in size-exclusion chromatography (SEC) (Table 1). MALS indicates that the I- and T-type proteins are probably present as a mixture of dimers and monomers in solution. The main peak (Fig. 5a) corresponds to the dimeric form, and this value is reported in Table 1. At high concentration (2–4 mg/ml), I- and T-type proteins are present as a mixture of oligomers. In contrast, monomeric and oligomeric fractions of C-type proteins YC_4A (Fig. 5a) and YC_8A (data not shown) can be separated, up to the highest concentration tested (4 mg/ml). However, the fractions of YC_4A and YC_8A, shown by MALS to be monomeric, elute earlier than expected for proteins of comparable size. The smaller YC_2A represents the only exception: independent of the concentration, the MALS-calculated mass values are always intermediate between monomer and dimer. A decrease in pH to 7 favors the formation of oligomeric species of I- and T-type proteins. C-type proteins are, in contrast, unaffected by pH (data not shown).

The circular dichroism (CD) spectra (Fig. 5b) indicate the presence of significant α-helical secondary structure content for all proteins, particularly for the I-type proteins. For I- and C-type consensus repeats, the absolute value of mean residue ellipticity (MRE) and the helical content generally increase

Table 1. Biophysical properties of designed and natural armadillo repeat proteins

Construct	Residues (repeats)[a]	p*I*[b]	MW_{calc} (kDa)[b]	Oligomeric state[c]	MW_{obs} (kDa)[d]	$MW_{obs/calc}$[e]	CD_{222} (MRE)[f]	Helical content (%)[g]	Observed T_m (°C)[h]
YI_2A	169 (4)	5.2	18.6	Dimer	64.6	1.7	−13,000	63	~55
YL_4A	253 (6)	4.8	27.4	Dimer	116.1	2.1	−19,500	80	~69
YI_8A	421 (10)	4.6	44.9	Dimer	148.8	1.7	−22,600	85	>85
YT_2A	169 (4)	6.3	18.6	Dimer	141.2	3.8	−7100	23	~56
YT_4A	253 (6)	6.5	27.3	Dimer	219.6	4.0	−10,100	40	~75
YT_8A	421 (10)	6.7	44.8	Dimer	229.7	2.6	−9400	35	~83
YC_2A	169 (4)	5.4	18.4	Mixture	59.1	n.d.	−9100	45	n.d.
YC_4A	253 (6)	5.1	26.9	Monomer	50.0	1.9	−12,100	49	n.d.
YC_8A	421 (10)	4.8	44.0	Monomer	76.7	1.7	−20,000	62	n.d.
YM_4A	253 (6)	5.1	27.1	Monomer	32.2	1.2	−18,800	87	~70
αArm[i]	435 (10)	5.5	48.2	Monomer	42.9	0.9	−14,300	54	~43
βArm[j]	528 (12)	8.7	57.6	Monomer	52.6	0.9	−16,800	60	~58

n.d. indicates that the value has not been determined due to either an inhomogeneous sample (oligomeric state of YC_2A) or lack of cooperative transition in thermal denaturation (YC_2A, YC_4A, and YC_8A).

[a] The number of residues includes the $MRGSH_6$ tag; the number of repeats includes capping repeats.

[b] p*I* and molecular weight calculated from the sequence; masses were confirmed by mass spectrometry.

[c] Oligomeric state as indicated by multiangle static light scattering.

[d] Observed molecular weight as determined in SEC.

[e] Ratio between observed and calculated molecular weight, taking into account the oligomeric state (Os): $MW_{obs/calc}=MW_{obs}/(Os\cdot MW_{calc})$.

[f] Mean residue ellipticity at 222 nm expressed as $deg\cdot cm^2/dmol$.

[g] Helical content estimated with the program CDpro.[41]

[h] T_m observed in thermal denaturation by CD.

[i] Armadillo domain of human importin-α1.

[j] Armadillo domain of mouse β-catenin.

with the number of internal repeats; in contrast, the helical content is almost constant for T-type proteins (Supplementary Fig. S4). The values of helical content were calculated using the program CDpro[41] and are indicated in Table 1.

The CD signal at 222 nm was chosen to monitor stability against thermal and denaturant-induced unfolding. I- and T-type proteins show a cooperative transition, while no transition was observed in C-type proteins (Fig. 5c). The midpoint of transition during thermal denaturation (T_m) increases with the number of repeats, for example, from approximately 70 °C for YL_4A to more than 80 °C for YI_8A (Table 1). Importin-α1 and β-catenin, containing 8 and 10

Fig. 5. Biophysical characterization of designed and natural armadillo repeat proteins. (a) SEC and MALS of designed armadillo repeat proteins containing four internal modules and of importin-α1. YL_4A, YT_4A, and YC_4A show apparent molecular weights higher than the globular proteins with the same calculated mass (about 27 kDa). The broad peaks shown by YL_4A and YT_4A are due to a mixture of dimers and monomers, as indicated by the molecular mass determined by light scattering. The highest point of the peak corresponds to the dimeric fraction. In the case of YC_4A, the first peak eluted contains probably a mixture of oligomers with high molecular masses. The monomeric peak after separation remains monomeric and was further characterized. The importin-α1 (αArm) is a monomer as indicated by LS and elutes at the expected volume. The data were obtained with a Superdex 200 column. The elution was followed by absorbance at 280 nm for YC_4A, YL_4A and αArm; YT_4A does not possess any residue absorbing significantly at 280 nm; thus, the elution was followed at 230 nm. V_0 indicates the void volume of the column. Alcohol dehydrogenase (ADH; MW=150 kDa), bovine serum albumin (BSA; MW=66 kDa), carbonic anhydrase (CA; MW=29 kDa), and aprotinin (Apr; MW=6.5 kDa) were used as molecular weight markers, and the corresponding elution volumes are indicated by the arrows. (b) CD spectra of I-type, T-type, and C-type proteins containing four internal modules. The natural armadillo domains of human importin-α1 (αArm) and mouse β-catenin (βArm) are indicated by open and filled circles, respectively. The values are reported as MRE. (c) Thermal denaturation curves. A comparison between designed armadillo repeat proteins containing four or eight internal modules is shown, from the top, for I-type, T-type, and C-type proteins. αArm and βArm are displayed in the bottom panel. The denaturation was followed by CD. The MRE at 222 nm is reported. (d) Thermal denaturation and renaturation of designed armadillo repeat proteins. From the top, YL_4A, YT_4A, and YC_4A are shown. For comparison, the bottom graph shows the irreversible denaturation of αArm. βArm shows a similar irreversible denaturation (data not shown). The denaturation was followed by CD. The values of MRE at 222 nm were normalized by setting the initial and the final values of the denaturation curves as 0 and 1, respectively. (e) Guanidinium-chloride-induced denaturation of armadillo repeat proteins containing eight internal modules. Comparison of YI_8A, YT_8A, and YC_8A with αArm. The denaturation was followed by CD. The values of MRE at 222 nm were normalized by setting the initial and the final values of the denaturation curves as 0 and 1, respectively. (f) Emission spectra of ANS in the presence of designed armadillo repeat proteins. YL_4A, YT_4A, and YC_4A are compared to αArm and βArm. I- and T-type proteins show fluorescence levels in the same range as natural proteins; in contrast, the fluorescence emission for C-type proteins is significantly higher and increases with the number of repeats. The values without buffer subtractions are shown. αArm was measured in a separate experiment and scaled according to the values of YC_4A present in both sets of experiments. Similar results were obtained with proteins containing two or eight internal repeats.

internal repeats, respectively, have lower midpoints of transition, even when compared with designed proteins with only 4 internal repeats (Table 1). It should be noted that the designed proteins retain a significant percentage of secondary structure at 95 °C and that the thermal unfolding is almost

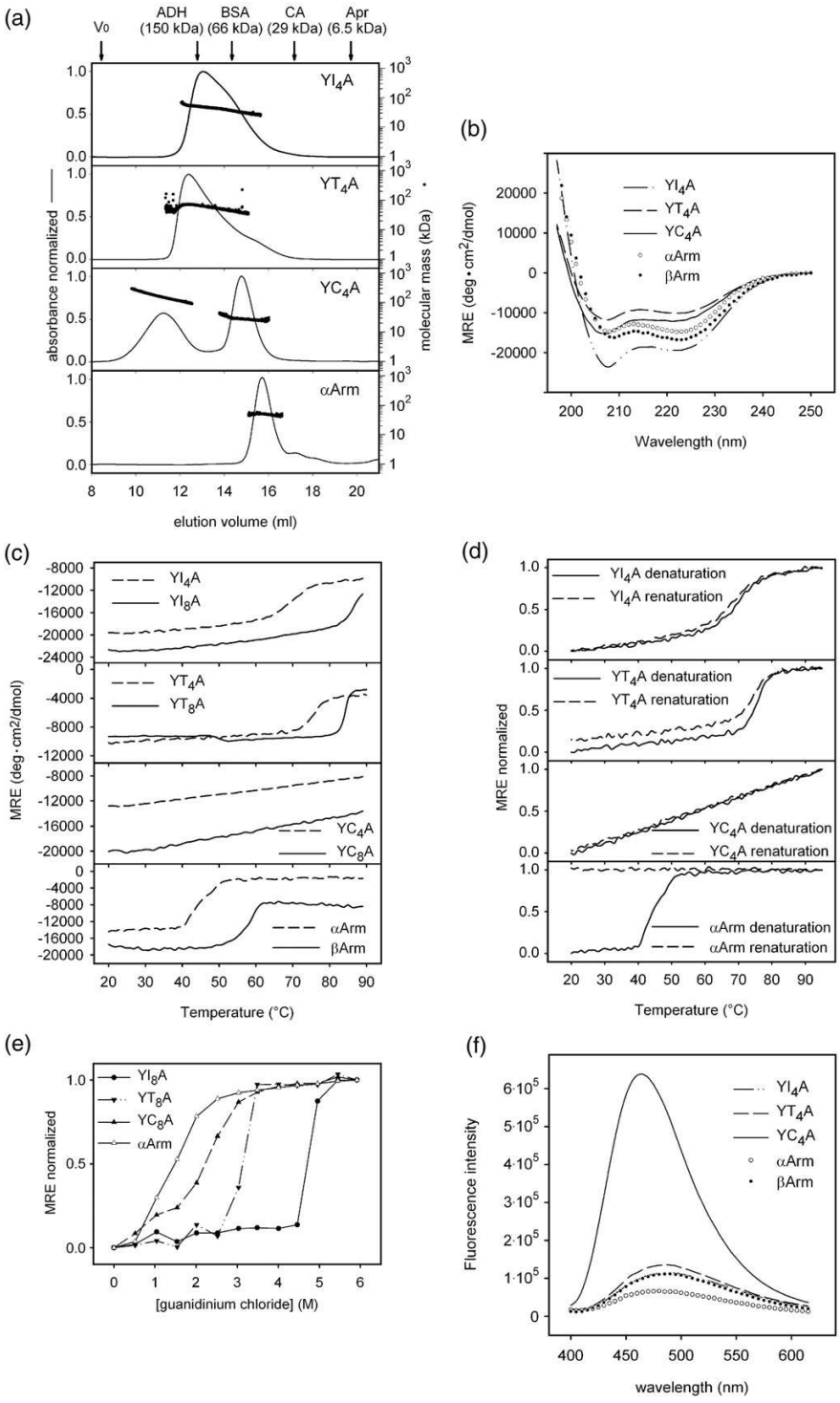


Fig. 5 (*legend on previous page*)

completely reversible, in contrast to natural armadillo proteins that cannot refold after thermal unfolding (Fig. 5d); YT_8A is the only designed armadillo repeat protein whose thermal unfolding is irreversible (data not shown).

We also investigated unfolding induced by guanidinium chloride. A direct comparison between natural and designed armadillo repeat proteins composed of 10 repeats (Fig. 5e) reveals for importin-α1 (αArm), with a midpoint of transition of 1.4 M guanidinium chloride, a lower stability than that for YI_8A and YT_8A, with approximately 4.8 and 3.2 M as midpoints of transition, respectively. YC_8A shows a gradual loss of secondary structure, especially at low concentrations of denaturant, apparently similar to αArm. Data from urea-induced unfolding experiments confirm the gradual loss of secondary structure for C-type proteins with increasing denaturant concentration. Natural armadillo domains show a stable pretransition baseline in unfolding induced by the weaker denaturant urea (data not shown).

The three types of consensus proteins (C-, I-, and T- type) also show a different behavior in 1-anilino-naphthalene-8-sulfonate (ANS) binding experiments. ANS is a fluorescent dye sensitive to the hydrophobic environment.[42] C-type proteins bind ANS strongly, suggesting the presence of an accessible hydrophobic core, while I- and T-type proteins show ANS binding in the same low range as the natural armadillo repeat proteins (Fig. 5f).

Thermal and guanidinium-induced denaturation and ANS results indicate that I- and T-type proteins share many characteristics with proteins with stable folds. Based on MALS data, however, I-type proteins are mainly present as dimers. T-type proteins show even higher deviations of elution behavior in SEC, and remarkably, the helical content does not seem to be significantly affected by the number of internal repeats, in contrast to the T_m value. C-type proteins, though monomeric, are characterized by strong ANS binding, an elution volume smaller than expected for a monomeric protein in SEC, and lack of cooperativity in thermal and chemical denaturation. These features, similar to some extent to the properties of molten globules,[43] indicate that the C-type proteins are probably not folded in a well-packed conformation, even though the expected secondary structure is detected by CD. Nonetheless, we chose the C-type proteins as the basis for our further investigations.

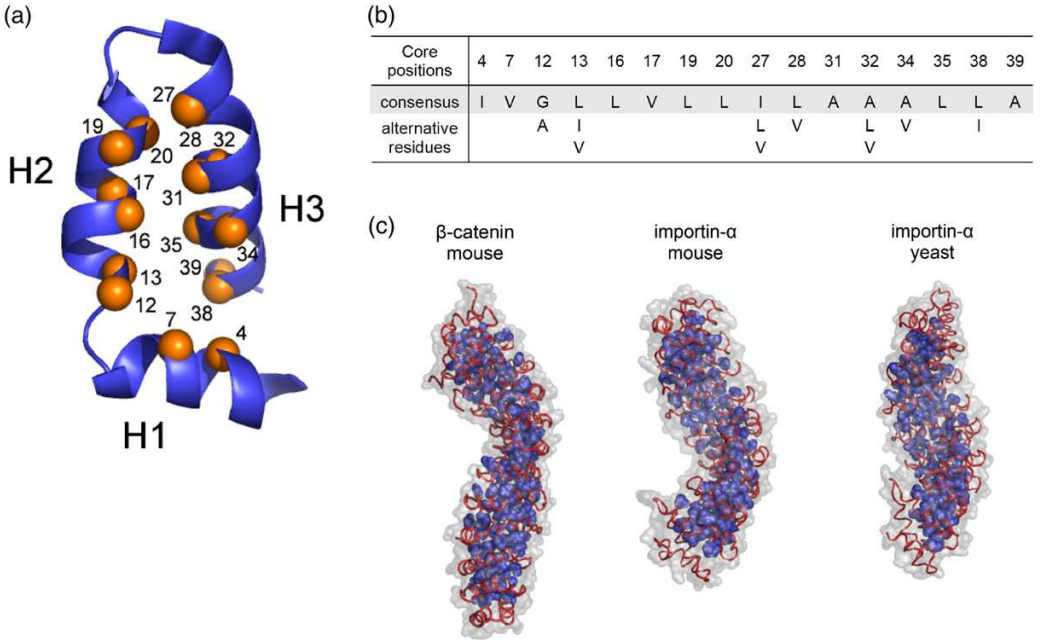
Consensus design improvement: Substitutions in the hydrophobic core

Due to the lack of conserved interrepeat hydrogen bonds and salt bridges, the tertiary structure of natural armadillo repeat proteins holds together mainly through nonpolar interactions. If the packing is not ideal, alternative conformations may become accessible. As a consequence, the molten-globule-like features of C-type proteins could be due to nonoptimal packing of the hydrophobic core.

The modular architecture of designed armadillo repeat proteins suggests that the computational search for a sequence leading to stable packing of the hydrophobic core might be achievable by considering a single repeat. However, the repeat can assume its correct conformation only in the context of a complete protein. It was, therefore, necessary to use the known structures of natural armadillo domains (comprising 400 to 500 residues) as templates for the sequence search.

The use of available algorithms (self-consistent mean field, dead-end elimination, genetic algorithm, and Monte Carlo search) for structures as large as armadillo domains has so far not been reported, despite recent achievements (reviewed by Butterfoss and Kuhlman[44]); such approaches would be, however, seriously compromised by the computational load and probably not even be possible in the case of dead-end elimination, as suggested by Voigt *et al.*[45] Therefore, we used here a different approach to treat a system of such size: information from sequence alignments was used to reduce the complexity in terms of variable positions and allowed residue types. The selected mutants were ranked according to energy values obtained by rotamer sampling. The method allows, in a simple way, to identify a number of hydrophobic core mutant sequences, which are likely to represent an improvement of the original C-type sequence.

The 16 positions contributing to the hydrophobic core in each repeat (Figs. 3 and 6a) were defined by having a solvent-accessible surface corresponding to less than 5% of the total residue surface, as determined by a probe with 1.4 Å radius. The final choice was made after visual inspection of the structures. The number of mutations was restricted to the most frequently occurring aliphatic amino acids at each position, based on the sequence alignment, while keeping the most conserved positions constant. Using these criteria, only 7 positions out of the 16 forming the hydrophobic core of a single repeat were allowed to vary and to host two or three different residue types (Fig. 6b). Mutants were modeled starting from three different backbones to average the influence of single structures out. Therefore, the structures of three different proteins {mouse β-catenin [Protein Data Bank (PDB) ID 2BCT],[22] yeast importin-α (PDB ID 1EE4),[46] and mouse importin-α (PDB ID 1Q1T[47])} were chosen to generate all the mutants (Fig. 6c). Model structures were constructed by substituting the core positions of every internal repeat with either the residues present in the C-type consensus or the aforementioned mutations (Fig. 6b). The initial rotamer conformations were randomly assigned. The noncore residues of the original structures were kept. In each structure, every repeat of the protein carries the same mutations. Structures corresponding to all the 432 combinations of allowed mutations, including also the set of residues of the original



Core positions	4	7	12	13	16	17	19	20	27	28	31	32	34	35	38	39
consensus	I	V	G	L	L	V	L	L	I	L	A	A	A	L	L	A
alternative residues			A	I V					L V	V		L V	V		I	

Fig. 6. (a) Hydrophobic core positions in a single repeat are indicated with orange spheres and the corresponding numbers. (b) Amino acids, in a single-letter code, allowed at the core positions during the calculations. The original amino acids present in the type C consensus are highlighted in gray. The total number of different combinations in each repeat is 432 ($2^4 \times 3^3$). The number of mutants is also 432 because the same mutation pattern was applied to all repeats in each protein. (c) Armadillo domains used as starting structures for the models of the mutants: murine β-catenin (PDB ID 2BCT) and importin-α from mouse (PDB ID 1Q1T) and *S. cerevisiae* (PDB ID 1EE4). The backbone trace is shown in red, and the protein surface is indicated in gray. The side chains belonging to the hydrophobic core residues, which correspond to the parts allowed to move freely during the simulation, are depicted in blue.

C-type consensus, were generated and subjected to energy minimization.

A sequence of heating–quench cycles (Fig. 7), followed by energy minimization, resulted in a series of structures and corresponding energy values that were used to generate the final ranking of the mutants (Supplementary Table S3). A detailed description of the rotamer sampling procedure is provided in Materials and Methods. Mutants with a hydrophobic core volume lower than the original consensus, calculated with values reported by Chothia,[48] were not included in the final ranking to reduce the number of false positives that might arise due to underpacking of the core (see Discussion).

Gene assembly, expression, and characterization of selected hydrophobic core mutants

Among the 30 top-ranked single repeat mutant sequences, 18 were selected for experimental validation. The best-ranking mutant sequence with low core volume was also selected to challenge the initial choice of a core volume filter during the ranking process (Table 2 and Supplementary Table S3). The influence of mutant repeats on the protein properties was experimentally evaluated in the format of proteins containing four identical internal repeats and Ny and Ca as capping repeats (Fig. 3). The original reference consensus sequence is thus YC_4A. The proteins were named with a progressive number, from mut1 to mut18; mut19 contains the sequence with low core volume.

The assembly of single repeats from oligonucleotides and the stepwise ligations were performed as described above, and the proteins were expressed and purified by IMAC in a single step with yields comparable to those obtained for YC_4A, that is, up to 100 mg/l of bacterial culture.

The experimental comparison was carried out by using CD, SEC, and binding of ANS. All the mutants share a similar CD spectrum with the original consensus but are characterized by a general increase in MRE at 222 nm, indicating a higher percentage of α-helical secondary structure. The increased elution volume of the mutants indicates a higher compactness of the proteins (Fig. 8) and correlates well with a decreased ANS binding. The mutant mut1, being a dimer, represents the only outlier, while all the other mutants are monomers, as indicated by MALS. Some of the core mutants carry additional mutations (indicated in Table 2), which were unintentionally introduced during the gene synthesis. Most of these mutations are located in the loops or at the surface of the helices and, thus, have probably only a small

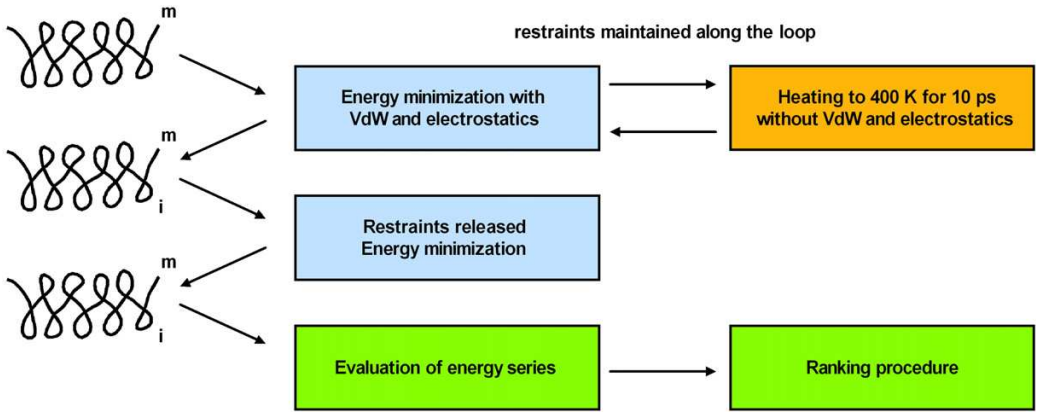


Fig. 7. Schematic diagram of the computational procedure for the evaluation of the hydrophobic core mutants. *m* indicates a particular mutant, and *i* is 1 of the 100 conformations of the mutant *m* obtained after each minimization step in the recursive sampling procedure. VdW, van der Waals interactions.

influence, if any, on the stability of the hydrophobic core; furthermore, they are present only in a single repeat out of four, reducing their overall contribution to protein properties.

Mutants mut2, mut3, mut4, mut7, mut11, mut12, and mut13 showed the best combination of low ANS binding and compactness, as judged by SEC, and were thus selected for further characterization by thermal denaturation. The mutant mut7 shows a significantly increased cooperativity during unfolding, compared to YC_4A and the other mutants (Supplementary Fig. S5).

The internal module corresponding to mut7, which was named M-type, contains three point mutations compared to the initial consensus sequence (Fig. 3). The mutant protein mut7, renamed YM_4A, is a stable monomer at several salt and protein concentrations, such as YC_4A; however, dimer formation of YM_4A was observed at pH 7 at high protein concentrations (5 mg/ml). No sign of precipitation or degradation was detected in protein solutions stored for up to 1 month at 4 °C in the IMAC elution buffer. The values for the biophysical properties examined are reported in Table 1.

The direct comparison of YC_4A and YM_4A is shown in Fig. 9. The [^{15}N,^{1}H]-heteronuclear single quantum coherence (HSQC) NMR spectra of YM_4A were recorded at pH 7, 8, 9, 10, and 11. YC_4A spectra

Table 2. Hydrophobic core of the selected mutants

	Hydrophobic core residues															
	4	7	**12**	**13**	16	17	19	20	**27**	**28**	31	**32**	**34**	35	**38**	39
C-type	I	V	**G**	**L**	L	V	L	L	**I**	**L**	A	**A**	**A**	L	**L**	A
mut1	–	–	**–**	**–**	–	–	–	–	**L**	**–**	–	**L**	**–**	–	**–**	–
mut2	–	–	**A**	**–**	–	–	–	–	**V**	**–**	–	**L**	**–**	–	**–**	–
mut3	–	–	**A**	–	–	–	–	–	**L**	–	–	**L**	–	–	–	–
mut4	–	–	**A**	–	–	–	–	–	**L**	–	–	–	–	–	**I**	–
mut5	–	–	**A**	–	–	–	–	–	**L**	–	–	–	**V**	–	–	–
mut6	–	–	**A**	**–**	–	–	–	–	**L**	**–**	–	**–**	**–**	–	**–**	–
mut7	–	–	**A**	–	–	–	–	–	**–**	–	–	**L**	–	–	**I**	–
mut8	–	–	**A**	**V**	–	–	–	–	**L**	–	–	**L**	–	–	–	–
mut9	–	–	**A**	**I**	–	–	–	–	**L**	–	–	**L**	–	–	–	–
mut10	**–**	**–**	**A**	**–**	**–**	**–**	**–**	**–**	**L**	**–**	**–**	**V**	**–**	**–**	**–**	**–**
mut11	–	–	**A**	**–**	–	–	–	–	**V**	**–**	–	**L**	**V**	–	**–**	–
mut12	–	–	**A**	–	–	–	–	–	**L**	–	–	**L**	**V**	–	–	–
mut13	–	–	**A**	**I**	–	–	–	–	**L**	–	–	**L**	**V**	–	–	–
mut14	**–**	**–**	**A**	**–**	**–**	**–**	**–**	**–**	**L**	**–**	**–**	**V**	**V**	**–**	**–**	**–**
mut15	–	–	**A**	**–**	–	–	–	–	**–**	**–**	–	**–**	**–**	–	**–**	–
mut16	–	–	**–**	**–**	–	–	–	–	**L**	**–**	–	**L**	**–**	–	**I**	–
mut17	–	–	**A**	**V**	–	–	–	–	**L**	–	–	**L**	**V**	–	**I**	–
mut18	**–**	**–**	**A**	**I**	**–**	**–**	**–**	**–**	**–**	**–**	**–**	**–**	**–**	**–**	**–**	**–**
mut19	–	–	**–**	**–**	–	–	–	–	**L**	**–**	–	**–**	**–**	–	**–**	–
I-type	–	–	**A**	**–**	–	–	–	–	**–**	**Q**	–	**L**	**–**	–	**I**	T

The numbers indicate the positions in the single repeat (cf., Fig. 3). The hydrophobic core positions subjected to mutation (12, 13, 27, 28, 32, 34, and 38) are indicated in boldface. The amino acids present at each position are reported as single-letter code. "–" indicates no difference with respect to C-type consensus. As a comparison, in the last row, the sequence corresponding to the I-type consensus is shown. An Ala→Thr mutation occurs in mut6 at position 15 in repeat 3, in mut9 at position 31 in repeat 4, in mut10 at position 12 in repeat 1, and in mut17 at position 15 in repeat 1. mut8 has a mutation Gly→Val at position 42 in repeat 4.

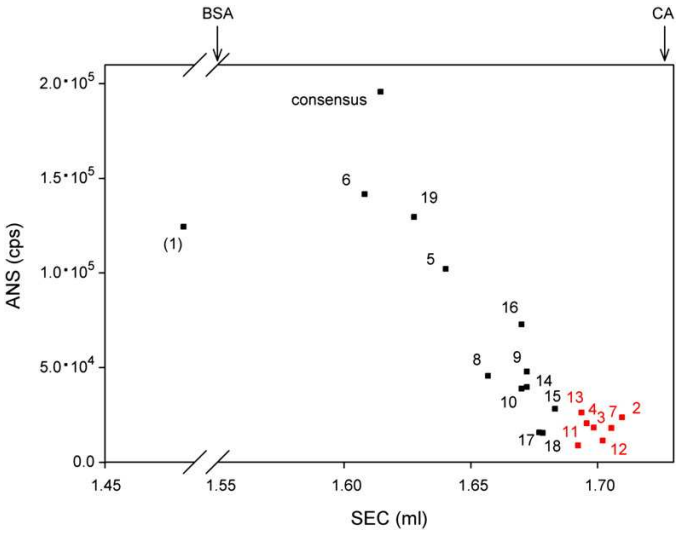


Fig. 8. Experimental evaluation of hydrophobic core mutants: elution volumes in SEC and fluorescence emission upon ANS binding. The numbers refer to the mutants reported in Table 1. *Consensus* indicates the protein containing four C-type internal repeats (YC_4A). All the proteins have a molecular mass of approximately 27 kDa. *mut1* (in parentheses) elutes before the consensus and the other mutants because of its dimeric state. All other mutants were shown to be monomeric by MALS. Peak values from absorbance at 280 nm in SEC and from fluorescence intensity are plotted. Errors in the measurements have been estimated with a subset of six proteins and two different preparations, leading to an average standard deviation of 0.01 ml for SEC and an average percentage error of 4% for ANS fluorescence intensity. As reference, carbonic anhydrase (CA; MW = 29 kDa) and bovine serum albumin (BSA; MW = 66 kDa) elute at 1.73 and 1.55 ml, respectively. The mutants depicted in red were selected for further characterization.

were collected at pH 6, 7, and 8. An increase in pH increases the line broadening of YC_4A but decreases it for YM_4A. Nevertheless, the overall dispersion is conserved for each protein at different pH values (data not shown). The YM_4A spectrum recorded at pH 11 and the YC_4A spectrum recorded at pH 6 are shown in Fig. 10. Amide proton frequencies of YC_4A are generally limited to the random-coil range (7.5–8.5 ppm), whereas many cross peaks of YM_4A are located outside this range. Moreover, the line widths from signals of YC_4A are slightly larger than those from signals of YM_4A. Increased line widths due to conformational exchange processes as well as limited signal dispersion are characteristic features of molten globule states of proteins.[49,50] Although no attempts have been made to assign the $^{15}N,^{1}H$ correlation map, $^{15}N\{^{1}H\}$-nuclear Overhauser enhancement (NOE) data were recorded to characterize internal backbone dynamics[51] and to probe for increased rigidity of YM_4A (data not shown). All detected amide moieties of YM_4A are characterized by $^{15}N\{^{1}H\}$-NOEs larger than 0.6, indicating well-folded segments, whereas for YC_4A, all the values are smaller than 0.3, many of which have negative NOEs, indicating a large flexibility. Thus, the NMR measurements confirm the molten-globule-like characteristics of YC_4A and the folded state properties of YM_4A.

Binding assay as functionality test

YM_4A and YC_4A share with natural importins a considerable number of residues critical for binding to nuclear localization sequences (NLSs), which are the natural ligands of importin-α proteins. Therefore, the designed proteins might retain some binding properties toward NLS. The NLS from the SV40 large T antigen[52] is considered a prototype sequence: it has been extensively studied in the literature and constitutes the reference point for the evaluation of NLS binding.[47]

The NLS from SV40 large T antigen (SPKKKRKVE) was expressed as a fusion protein with phage lambda protein D (pD), biotinylated, and immobilized on NeutrAvidin-coated plates. Being of similar size, the hemagglutinin tag (YPYDVPDYA, here referred to as HA), also fused to protein D, was used as a negative control. ELISA experiments (Fig. 11) reveal that both YM_4A and YC_4A bind specifically to the NLS and that the binding can be competed by a free NLS peptide in solution. However, the unspecific binding of YM_4A to HA and NeutrAvidin is reduced in comparison to YC_4A.

In summary, even though the high concentrations of protein and competing peptide indicate a rather weak affinity, YM_4A was able to specifically recognize the same target as the natural armadillo repeat proteins and to reduce the unspecific binding observed for YC_4A, further validating the design process.

Discussion

Consensus design

Consensus design has been successfully applied in this work to generate designed armadillo repeat proteins. Similar to leucine-rich repeat proteins,[30] but in contrast to ankyrin repeat proteins[27] and tetratricopeptide repeats,[29] different subfamilies can be clearly defined in the case of armadillo repeat proteins, based on sequences and available structures. Out of 42 signature positions, 12 are char-

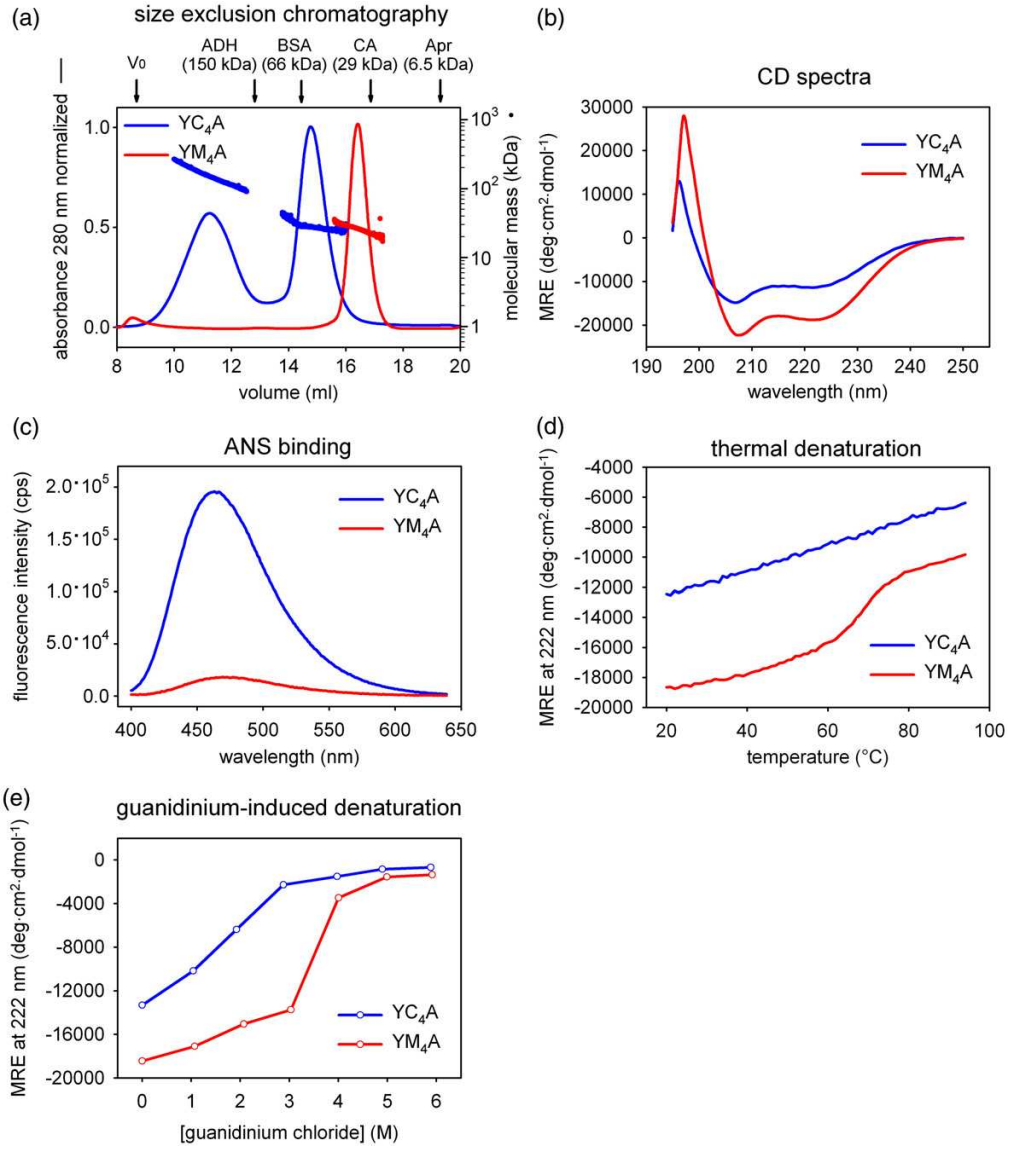


Fig. 9. Comparison between YC_4A, in blue, and YM_4A, in red. SEC (a) was performed with samples directly after IMAC purification. MALS data are also shown. The chromatogram of YC_4A displays one peak corresponding to the monomer (on the right) and one corresponding to oligomeric fractions (on the left). CD spectroscopy (b) shows an increase in ellipticity for YM_4A. ANS binding (c) is drastically reduced for YM_4A to levels typical of natural armadillo repeat proteins; the data shown refer to values after buffer subtraction. Thermal denaturation (d) and guanidinium-induced denaturation (e) indicate the presence of a cooperative unfolding transition, characteristic for native-like proteins, for YM_4A. V_0 indicates the column void volume. Alcohol dehydrogenase (ADH; MW = 150 kDa), bovine serum albumin (BSA; MW = 66 kDa), carbonic anhydrase (CA; MW = 29 kDa), and aprotinin (Apr; MW = 6.5 kDa) were used as molecular weight markers, and the corresponding elution volumes are indicated by arrows.

acteristic for armadillo repeats, but the conservation at other positions is relatively low.[32] To obtain a more reliable and informative consensus, we deemed it necessary to analyze the subfamilies independently. The use of closely related sequences should also improve the self-compatibility between designed repeats. At the time of the initial sequence design, only members of importin-α and β-catenin/plakoglobin subfamilies were known to be peptide binders and had crystal structures available. As a consequence, only the repeats from proteins belonging to these subfamilies were thus chosen for the calculation of the consensus, to avoid interference from other subfamilies of unknown structure that could negatively affect the final sequences. Indeed, the later publication of the structure of plakophilin[53] (a member of the p120 subfamily) revealed an unexpected shape with a pronounced bend in the

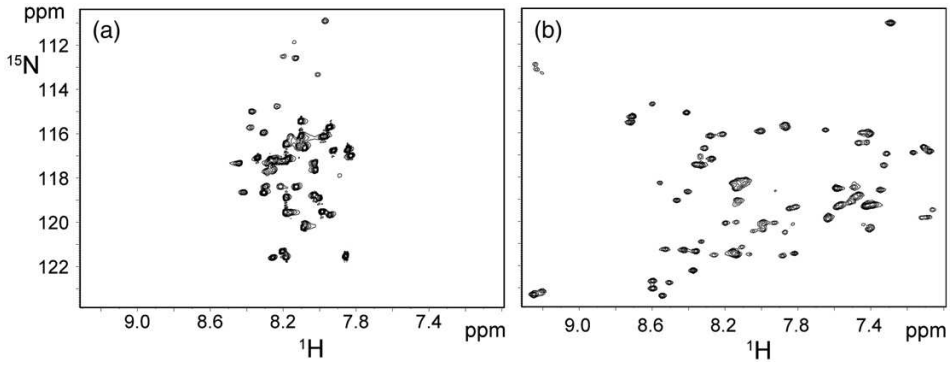


Fig. 10. [^{15}N,^{1}H]-HSQC spectra of designed armadillo repeat proteins: YC_4A (a) at pH 6 and YM_4A (b) at pH 11. Both spectra were recorded at a temperature of 310 K in 20 mM Tris–HCl and 30 mM NaCl. The protein concentration was 0.6 mM.

middle of the domain, supporting, *a posteriori*, the initial choice of sequence restriction to the subfamilies mentioned above.

An overall consensus was, however, also realized to take into account the possible combination of sequences belonging to importin-α and β-catenin/plakoglobin subfamilies. An obvious concern regarding the combination of these two subfamilies was that the overall consensus (type C) might be too similar to the importin consensus (type I) due to the slight overrepresentation of importin sequences in the original set. After the exclusion of the positions involved in binding, highly conserved for functional reasons especially in the importin subfamily and thus preserved also in the overall consensus, the C- and I-type repeats share 74% identity and 82% similarity, while C- and T-types have corresponding values of 70% identity and 87% similarity. The values indicate that the overall consensus is thus not significantly biased toward the importin consensus in the "framework" positions, that is, the positions not responsible for binding. The positions involved in peptide binding will be randomized in the library design and thus do not play a role in these considerations. Nevertheless, despite the similarity between I-, T-, and C-type modules, we always used only one type of consensus modules in every repeat protein tested, to provide a constant interface between the repeats and to be able to correlate the protein properties with the types of modules.

The capping repeats represented a second key point in the protein design. As observed for designed ankyrin repeat proteins,[31,54] capping repeats can dramatically increase *in vivo* folding yield and prevent aggregation. We found that an N-terminal capping repeat derived from yeast importin-α (Ny) and an artificial C-terminal capping repeat (Ca), designed by replacing exposed hydrophobic residues, give the highest expression yield of soluble protein in *E. coli*. Remarkably, we could find a single combination of capping repeats that allowed us to analyze the properties connected to the types of internal modules.

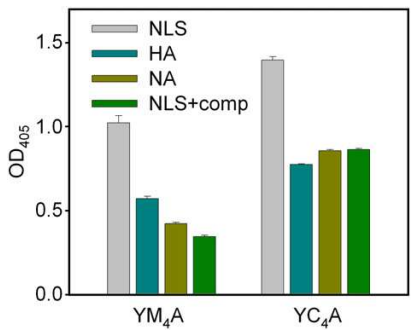


Fig. 11. ELISA of YM_4A and YC_4A. YM_4A binds specifically to immobilized SV40 large T antigen NLS. Immobilized hemagglutinin tag peptide (HA) and NeutrAvidin (NA) are negative controls. Binding to NLS can be competed by addition of NLS peptide (SPKKKRKVE) in solution at a concentration of 10 μM (NLS+comp). Experiments were performed in duplicate with YM_4A and YC_4A at a concentration of 1 μM.

Protein properties

Data from the artificial repeat proteins previously designed[28–30] indicate that biophysical properties often correlate with the number of internal repeats. Indeed, this behavior was also observed for designed armadillo repeat proteins.

I-type proteins

I-type proteins show that helical secondary structure content, thermal stability, and resistance to guanidinium-induced denaturation increase with the number of repeats, pointing in the same direction as data from other artificial repeat proteins.[27,29,30,55] A helical content of approximately 80% for YI_4A and YI_8A (Table 1) corresponds to the expected theoretical value from the design and is even higher than the values observed for natural armadillo domains. Low ANS binding and clearly defined transitions in thermal and guanidinium-

induced denaturation indicate that the I-type module can lead to native-like molecules, and the elevated midpoint of denaturation points toward a superior thermodynamic stability compared to natural proteins. At the same time, thermal denaturation is almost completely reversible. The thermal denaturation was employed here as a qualitative method to assess the stability of the designed proteins and to compare them to their natural counterparts. A detailed thermodynamic analysis requires, however, further investigation of the folding mechanism, which is probably more complex than a simple two-state transition and possibly also described by the Ising model as in the case of other designed repeat proteins.[56,57]

I-type proteins could, thus, be good candidates as scaffold for peptide-binding molecules. However, their predominant dimeric state constitutes a disadvantage during selection and characterization of binding properties due to possible avidity effects. Even considering that the I-type proteins are dimers, the SEC data indicate an elution volume still larger than expected, which could be interpreted as a result of an elongated shape. It is noteworthy that natural armadillo proteins do not show a higher-than-expected apparent mass in gel filtration (Table 1).

T-type proteins

T-type proteins share several native-like characteristics with I-type proteins, such as the presence of a compact hydrophobic core inaccessible to solvent, as suggested by ANS binding levels that are as low as those of natural armadillo repeat proteins, and the transitions observed in thermal and guanidinium-induced denaturation. The reversibility of thermal denaturation in T-type proteins is, however, less complete than that in I-type proteins and completely lost in the case of YT_8A. The helical content of T-type proteins is approximately independent of the number of repeats and generally lower than that in natural armadillo repeat proteins. The gel filtration results are, however, similar to I-type proteins, with apparent molecular masses higher than expected by more than a factor of 3 on average, already taking the dimeric state into account (Table 1). Due to the native-like properties of T-type proteins, the increase in hydrodynamic radius could be interpreted as an effect of a rodlike shape. Despite the different behavior in gel filtration, T-type proteins are therefore more similar in their biophysical characteristics to native armadillo proteins than to an idealized scaffold. When applying a strategy of protein assembly from preselected modules, which represents one of the aims of a general modular peptide binder, the scaffold properties should ideally change in a regular and predictable way, when adding modules, without altering the general characteristics. However, this is not observed for YT_8A, where the reversibility in thermal unfolding is completely lost.

C-type proteins

C-type proteins, in contrast to the other designed armadillo proteins, do not show a clear transition in thermal or guanidinium-induced denaturation but a gradual loss of secondary structure, and ANS binding results indicate the presence of an accessible hydrophobic core. Thus, C-type proteins are probably not completely folded but rather are in a molten-globule-like state. The secondary structure is present, as indicated by CD, but the proteins retain a high level of flexibility due to the lack of a fixed tertiary structure. The high apparent molecular masses observed in gel filtration, where MALS indicate monomeric states, might then be interpreted as a consequence of the intrinsic flexibility of the polypeptide chain. The molten-globule-like characteristics of C-type proteins represent a serious limitation in library generation, where framework stability and tolerance to mutations are desired. From the point of view of the design, however, the observation of a molten-globule-like state for armadillo repeat proteins built from overall consensus (type C) modules could suggest either an insufficient stability of each repeat or an inadequate interaction between them, supporting the initial choice of restricting the design to specific subfamilies.

Molten globule stabilization

The initial consensus-based approach led to stable dimers (I- and T-types) or molten-globule-like monomeric proteins (C-type). The further possible design steps to obtain a stable monomer were either the disruption of the interaction in the dimer or the stabilization of the C-type proteins. However, no information was available concerning the dimerization interface and the residues involved; both surface interaction and domain swap were conceivable as mechanisms of dimer formation. The improvement strategy would thus have to involve systematic point mutations of several single residues and combination of residues, with the risk that the disruption of the dimer will simply lead to a stability loss or even a molten globule state. For disrupting a dimer, the improvement strategy would have to consist of systematic mutations of single residues or combination of residues without a structural hint to select mutations.

We chose instead an alternative approach, focused on the stabilization of the hydrophobic core of the molten-globule-like C-type proteins, using a computational approach. As mentioned above, the molten-globule-like state suggested inadequate inter- or intra-repeat interactions at the hydrophobic core level and, hence, insufficient packing of the core. Our results show that the introduction of only three point mutations in the hydrophobic core of C-type repeats was sufficient to convert a molten-globule-like protein with four internal repeats to a stable conformation. This strongly argues that the packing of the hydrophobic core was indeed the critical parameter for obtaining a stable fold.

The underpacking of the core may be one of the reasons for the molten globule behavior[58] of the C-type proteins. Two of the mutations (Gly to Ala at position 12 of the repeat and Ala to Leu at position 32) increase the calculated volume of the hydrophobic core, bringing it close to the average value of natural repeats. These residues are also the most common among the 50 highest-ranking sequences with a frequency of 72% for Ala12 and 50% for Leu32. The third mutation (Leu to Ile at position 38) probably reduces the local flexibility by limiting the number of available rotamers. Such a restriction can help to lock the hydrophobic core in a unique conformation, and this positive effect could overcome the disadvantage of having a residue with low helical propensity, such as isoleucine. However, as observed for several mutants, the contributions from the single residues are not additive and the core packing is the result of a particular combination of residues.

Strikingly, the M-type repeat has, among the 432 mutants screened *in silico*, the core sequence closest to the I-type repeat (Table 2). The only two core residues that differ between M- and I-type repeats (Gln28 and Thr39) were not included in the set of possible mutations in our computational approach. The protein YI_4A, derived from I-type modules, shows characteristics very similar to YM_4A, apart from its dimeric state. Therefore, the particular core sequence obtained for both types of repeats represents a reliable solution for core packing, considering that it has been obtained by consensus design (for I-type) and simulation (for M-type). The hydrophobic core is probably rather stable, and we may speculate that the dimerization observed for I-type proteins takes place most likely via surface interaction instead of domain swap. The introduction of surface point mutations could then possibly lead to the formation of stable monomers.

YM_4A

The observed biophysical characteristics indicate that YM_4A represents a significant improvement of the original consensus sequence. YM_4A is almost as compact as globular proteins with similar molecular weight, as judged from elution volumes in SEC, and only marginally binds ANS, with values in the range observed for natural armadillo repeat proteins. The thermal and guanidinium-induced denaturation curves have sigmoidal profiles, indicating the presence of a cooperative unfolding, a hallmark of natural globular proteins.

NMR

NMR spectra provide further indications of the folded structure of YM_4A. Due to the repetitive nature of the sequence, it is *a priori* not clear how many peaks should be expected, but, even in the absence of specific assignments and considering the effects of symmetry, most of the peaks are probably present. However, Gly residues, usually observed in a characteristic region of the correlation map, are missing, most likely due to the highly accelerated amide proton exchange at pH 11 for residues outside the regions of secondary structure. Nevertheless, the presence of most peaks at the elevated pH indicates that the majority of amide moieties are protected from solvent access.

Although it was not possible to assign the spectra, the $^{15}N\{^1H\}$-NOE data indicate that almost all peaks in the proton–nitrogen correlation spectrum correspond to amide moieties with motional properties similar to those of residues from stably folded secondary structural elements. Hence, the NMR measurements suggest that YM_4A at pH 11 can be considered as a well-folded globular protein, whereas YC_4A shows characteristics of a molten globule. YM_4A, at pH lower than 10, displays broader lines, without affecting the signal dispersion, indicating that under those conditions, good side-chain packing is probably disturbed by the presence of an ionized group. A large range of pH values was also tested for YC_4A, yet without leading to any improvement in the dispersion of the signals in $[^{15}N,^1H]$-HSQC spectra or narrowing of the line width. Hence, electrostatic interactions are not dominating the molten globule properties of YC_4A. These observations rather indicate that subtle effects of side-chain packing are involved and that proper side-chain packing is achieved only in YM_4A, which presumably requires a neutral state of one group that is charged at neutral pH but uncharged at pH 11. As the lines are sharper at basic pH, lysine residues are the candidates for causing this effect, because of the possible repulsive interaction with the lysines in the neighboring repeats when both are charged, as observed in the molecular models.

Peptide binding

The binding to the SV40 large T antigen NLS observed in ELISA confirms the interpretation of the biophysical data. Unspecific binding has been reduced in YM_4A, compared to the original molten-globule-like YC_4A, as observed in binding to NeutrAvidin and to the hemagglutinin tag and in the competition experiment.

Even though no design effort was made in the present work for binding to a target peptide, YM_4A and YC_4A do show a weak but specific binding, indicating a correct disposition of the residues involved in interactions with the peptide. Glu30, Trp33, and Asn37 in the M-type repeat correspond to the residues responsible for binding to NLS in natural importin-α proteins. These residues are present in YC_4A and YM_4A due to the high conservation in the importin-α sequences, which were used in the original consensus design. The competition with soluble peptide strongly suggests the presence of specific interactions rather than a merely electrostatic binding phenomenon.

Further experiments will be needed to clarify the binding of consensus-designed armadillo repeat proteins. Nevertheless, the results already achieved indicate a correct structure. Armadillo repeat proteins based on M-type repeats can thus be used as scaffold for library generation and selection.

Evaluation of the computational method

A rotamer sampling method was chosen to identify, from a large pool of candidates, armadillo repeat proteins with improved core packing. The approach was devised for use with large proteins, up to 500 residues in our case. Despite recent advances,[59] such complexity is still not easily treatable by the available methods for core repacking, which proceed through a cycle of mutation, selection of residue conformation, and energy minimization. In terms of computational load, the search for a sequence with minimal energy is highly demanding, or even not affordable at all, for large proteins. In contrast, a simple evaluation of the potential energy of protein models after energy minimization is rather unreliable. The introduction of point mutations in the hydrophobic core requires the rearrangement of the core side chains to optimize the core packing. This task is, however, not fulfilled by a simple energy minimization, especially when the energy barrier between rotamers is too high to be overcome and only the nearest local minima for the side chains are reached (e.g., for tightly packed side chains). However, our aim was not to find the conformation at the global minimum but to estimate the packing efficiency of given mutants. A random sampling, helped by the partial removal of the energy barriers and followed by statistical analysis, is thus a feasible procedure for evaluating the packing of each mutant protein.

Though being a simplified approach, it was still necessary to reduce the complexity of the system. The choice of candidate sequences was based on information derived from sequence alignment and interactions in the structures. This approach is not exhaustive, but it restricts the search space to the most promising mutants according to criteria that are independent of the computational method.

Nevertheless, some further restrictions were required to keep the computational load within reasonable boundaries. Calculations without solvation terms are computationally less expensive and can be applied in our case, where the mutations correspond only to aliphatic–aliphatic substitutions in the hydrophobic core and are not expected to influence solute–solvent energy contributions. When restraining backbone atoms, the influence of the core mutations on the surface electrostatics is negligible and the electrostatic contribution to the potential energy does not vary upon mutation.

A second crucial issue was the choice of three different armadillo structures as starting points for the generation of the mutant models to avoid a result biased by the use of a single structure. Additionally, the introduction of multiple backbone templates and backbone flexibility has already been shown to improve the quality of the sequence search.[60,61] In the present work, we did not attempt to build a new backbone including the information from multiple structures, but we allowed "flexibility" by using three starting structures and fixing the coordinates of the backbone atoms using harmonic restraints. The use of three structures per mutant also helped to enhance the signal-to-noise ratio of the ranking, as a high rank for all three armadillo structural backgrounds is generally required to obtain a high overall rank.

Sequences with low core volume were excluded from the final ranking to decrease the number of false positives in the pool selected for experimental characterization (Supplementary Table S3). One example of such a low core volume sequence is the original C-type consensus, which is highly ranked; its core volume was set as the lower volume threshold for the discrimination of the mutants. On a fixed backbone, a reduced core volume allows side chains to assume nearly ideal values of bond lengths, angles, and dihedrals, leading to a significant reduction of the total energy and to an artificially high rank. An increase in flexibility of the backbone could also be the source of artifacts: the cavities in the hydrophobic core of low volume mutants can be compensated by compressing the backbone structure and bringing the side chains close enough to take advantage of the van der Waals interactions. However, a reduction of the backbone flexibility could be detrimental for mutants slightly more overpacked than the natural structures, which would not be able to reach low energy values without backbone adjustments.

Homology models of C-type proteins, based on the armadillo crystal structures, indicated the likely presence of small cavities in the hydrophobic core, suggesting that underpacking is one of the possible reasons for the molten globule state. It is thus unlikely that proteins with core volume lower than the original C-type consensus can provide better packing. Furthermore, mut19, the highest-ranked mutant with low core volume, displays ANS binding and SEC properties closest to the original overall consensus sequence, confirming the validity of our selection filter based on core volume. No threshold level was set for possible overpacking cases: the maximal value of core volume among the considered mutants was still in the range of the average repeat volume calculated for the reference structure of murine importin-α (PDB ID 1Q1T; Supplementary Table S3).

On the experimental side, the use of ANS binding and SEC to discriminate mutants is rather qualitative but can represent an efficient and relatively fast method for screening. A good overall indication of the quality of our method is given by the fact that all the monomeric mutants analyzed show an improvement compared to the original overall consensus.

Conclusions

This work focused on the generation of designed armadillo repeat proteins for the construction of a general modular peptide-binding scaffold. An initial consensus-based design led to well-expressed and stable but dimeric proteins or molten globules. A stable, well-expressed monomeric protein was obtained using a force field-based approach for the stabilization of the hydrophobic core of the molten globule variant.

In a library perspective, a monomeric protein allows a better evaluation of the binding properties, without the influence of possible avidity effects, which can be critical in the discrimination between similar target peptides. The mutations to be introduced to generate a library will only affect surface residues, leaving the hydrophobic core untouched except for one position (position 4 may contribute to both the hydrophobic core and the binding site). Therefore, the favorable characteristics of the designed proteins will probably be kept for most library members and selected specific binders.

Materials and Methods

Sequence analysis and modeling

SMART†,[33,34] Swiss-Prot‡,[26] and PDB§[62] were used as the starting databases for our analysis. GCG (Wisconsin Package Version 10.3, Accelrys Inc., San Diego, CA), BLAST‖,[63,64] and ClustalW¶[35] were used for sequence retrieval and alignment. Structure analysis and modeling were performed with Swiss-Pdb Viewer[a],[65] MOLMOL[b],[66] PyMOL[c] (DeLano Scientific LLC, San Francisco, SA), and INSIGHT II (Accelrys Inc.). Vector NTI (Invitrogen) was used for vector and oligonucleotide design.

General molecular biology methods

Unless stated otherwise, experiments were performed according to Sambrook and Russell.[67] Vent Polymerase (New England Biolabs, USA) was used for all DNA amplifications. Enzymes and buffers were from New England Biolabs or Fermentas (Lithuania). The cloning and production strain was *E. coli* XL1-blue (Stratagene, USA). Competent cells were prepared according to Inoue *et al.*[68] The *E. coli* strain M15 (Qiagen, Germany), containing the plasmid pREP4, was used for the production of ^{15}N-labeled proteins for NMR experiments. The cloning and protein expression vectors were pQE30 (Qiagen, Switzerland) and pPANK (GenBank accession number AY327140). From this, the vector pPANK-NyCa was constructed by cloning of the capping repeats Ny and Ca. pPANK-NyCa contains the BsaI and BpiI restriction sites between the capping repeats for cloning purposes. Note that the inserts were constructed with a double stop codon (see Supplementary Fig. S3). pPANK-NyCa was used to clone the internal repeats for N8C and core mutant proteins. pQE30 and derivatives such as pPANK carry an $MRGSH_6$ tag at the N-terminus of the proteins. The DNA sequences corresponding to the NLS and HA peptides were inserted in the vector pAT223 (GenBank accession number AY327138) and expressed as fusion proteins to pD. The produced proteins consist of N-terminal Avi tag, pD, His_6 tag, and the peptide at the C-terminus. The plasmid pBirAcm (Avidity, USA), encoding *E. coli* biotin-protein ligase BirA, was used for *in vivo* biotinylation of pD peptides.

Cloning of designed armadillo repeat proteins

Oligonucleotides were purchased from Microsynth AG (Balgach, Switzerland). A complete list of all oligonucleotides used is given in Supplementary Table S1. An approach similar to the one described by Binz *et al.*[27] was adopted for gene assembly (Supplementary Fig. S2). All single repeat modules were assembled from oligonucleotides by assembly PCR. The single modules of the core mutants were assembled using the combinations of oligonucleotides indicated in Supplementary Table S2. As an example, for the C-type consensus, pairs of partially overlapping oligonucleotides (1–2, 3–4, and 5–6) were annealed and the double strand was completed by PCR. Then, 2 μl from these PCR reaction mixtures was combined as template for a second assembly reaction in the presence of oligonucleotides 1 and 6. All the oligonucleotides were used at a final concentration of 1 μM. The annealing temperature was 47 °C for the first reaction and 50 °C for the second. Thirty PCR cycles were performed with an extension time of 30 s. The same procedure was applied for the other internal and capping repeats. Only four oligonucleotides were used for the N-terminal capping repeats. BamHI and KpnI restriction sites were used for the direct insertion of the modules into the plasmid pQE30. The single modules were PCR amplified from the vectors, using external primers pQE_f_1 and pQE_r_1 (Qiagen, Switzerland). Neighboring modules were digested with the type IIS restriction enzymes BpiI and BsaI and directly ligated together. The genes coding for the whole proteins were assembled by stepwise ligation of the internal and capping modules. BamHI and KpnI restriction sites were used for insertion of the whole genes into the vector pQE30 and the plasmids were sequenced. For pD-peptide fusions, oligonucleotides encoding both strands of the peptide sequences and containing the restriction sites for BamHI and HindIII were mixed and heated to 95 °C for 10 min and then cooled to 4 °C to allow annealing of the two strands. The double-stranded DNA fragments were subsequently digested with BamHI and HindIII and ligated into the plasmid pAT223.

Natural armadillo domain constructs

The armadillo domain of mouse β-catenin (βArm; residues 150–665) was amplified from the cloned β-catenin gene[22] (a generous gift from W.I. Weis, Stanford University, USA) using oligonucleotides AcatFOR and AcatREV, digested with BamHI and KpnI, and inserted into pQE30. The armadillo domain of human importin-α1 (αArm;

‡ http://www.expasy.org
§ http://www.pdb.org
‖ http://www.ncbi.nlm.nih.gov/blast
¶ http://www.ebi.ac.uk/clustalw
[a] http://www.expasy.org/spdbv/
[b] http://hugin.ethz.ch/wuthrich/software/molmol/
[c] http://pymol.sourceforge.net

residues 83–505) was amplified from a vector containing the importin gene (named importin-α5 in the original publication,[69] a generous gift from M. Köhler, Ostseeklinik Damp, Germany) using oligonucleotides IMAF5 and IMAR5, digested with BamHI and KpnI, and inserted into pQE30. Both proteins carry an N-terminal $MRGSH_6$ tag, as do the designed armadillo repeat proteins.

Protein expression and purification

E. coli XL1-blue cells were transformed with the respective plasmid and grown in LB medium containing 1% (w/v) glucose and 50 μg/ml of ampicillin at 37 °C with vigorous shaking. Expression was induced by IPTG (final concentration of 0.5 mM) when the culture reached OD_{600}=0.6. After 3 h of expression, cells were harvested by centrifugation. For *in vivo* biotinylation of pD peptides that contain an N-terminal Avi tag, cells were cotransformed with pBirAcm and pAT223 (carrying the pD-peptide constructs) and grown in medium containing 30 μg/ml of chloramphenicol and 50 μg/ml of ampicillin. Before induction with IPTG, biotin was added to the medium to a final concentration of 50 μM, according to Cull and Schatz.[70]

Protein purification was performed at 4 °C. Cells were resuspended in 50 mM Tris–HCl and 500 mM NaCl (pH 8.0) and lysed in a French pressure cell (SLM Instruments, USA) at a pressure of 1200 psi. The lysis mixture was further homogenized by sonication (Branson, USA). Insoluble material was pelleted by centrifugation at 20,000*g* for 30 min. The supernatant was purified by IMAC with Ni–NTA material (Qiagen), equilibrated with buffer containing 50 mM Tris–HCl, 500 mM NaCl, 10% (v/v) glycerol, and 20 mM imidazole (pH 8.0). Columns were washed extensively with the equilibration buffer and then proteins were eluted with an elution buffer identical with the equilibration buffer but also containing 250 mM imidazole. β-Catenin was expressed and purified under the same conditions.

For importin-α1, the expression was carried out at 25 °C for 6 h and the cell pellet was resuspended in lysis buffer containing 50 mM Tris–HCl, 500 mM NaCl, 10% glycerol, 5 mM β-mercaptoethanol, and 10 mM imidazole (pH 8.0). IMAC purification was performed as indicated above using the same buffers with the addition of 5 mM β-mercaptoethanol. Samples were then dialyzed overnight against 50 mM Tris–HCl and 2 mM DTT (pH 8.0) and applied to a POROS HQ anion-exchange column, equilibrated with running buffer (50 mM Tris–HCl, pH 8.0), using the BioCAD 700 E Perfusion Chromatography Workstation (Applied Biosystems, Germany). The column was then washed with 50 mM Tris–HCl and 20 mM NaCl (pH 8.0), and the samples eluted with a gradient from 20 mM to 1 M NaCl. Protein size and purity were assessed by 15% SDS-PAGE, stained with Coomassie PhastGel Blue R-350 (GE Healthcare, Switzerland).

The expected mass of all the studied proteins was confirmed by mass spectrometry. Protein concentrations were determined by absorbance at 235 and 280 nm using molecular masses and extinction coefficients calculated with the tools available at the ExPASy proteomics server‡ and by the bicinchoninic acid assay (Pierce).

SEC and MALS

Analytical SEC was carried out either on an Ettan LC system using a Superdex 200 PC 3.2/30 column (flow rate 70 μl/min) or on an ÄKTA explorer chromatography system using a Superdex 200 10/30 GL column (flow rate, 0.5 ml/min) (GE Healthcare). Phosphate buffer (50 mM phosphate and 150 mM NaCl, pH 7.4) and two Tris-based buffers (20 mM Tris–HCl and 50 mM NaCl, pH 8.0, or 50 mM Tris–HCl and 500 mM NaCl, pH 8.0) were used. The armadillo domain of β-catenin was soluble only at 150 or 500 mM salt concentration. The armadillo domain of importin-α1 was analyzed in phosphate buffer (50 mM phosphate, 500 mM NaCl, and 5 mM DTT, pH 7.4). The core mutants were analyzed in buffer containing 20 mM Tris–HCl and 50 mM NaCl, pH 8.0. MALS measurements were performed with a miniDAWN light-scattering detector and an Optilab refractometer (Wyatt Technologies, USA) coupled to the ÄKTA system. Molecular weight estimates were calculated using the ASTRA 4.73.04 software package (Wyatt Technologies).

CD spectroscopy

CD measurements were performed on a Jasco J-810 spectropolarimeter (Jasco, Japan) using a 0.5-mm cylindrical thermocuvette. CD spectra were recorded from 190 to 250 nm with a data pitch of 1 nm, a scan speed of 20 nm/min, a response time of 4 s, and a bandwidth of 1 nm. Each spectrum was recorded three times and averaged. Measurements were performed at 20 °C. The CD signal was corrected by buffer subtraction and converted to MRE. Heat denaturation curves were obtained by measuring the CD signal at 222 nm with temperature increasing from 20 to 95 °C (data pitch, 1 nm; heating rate, 1 °C/min; response time, 4 s; bandwidth, 1 nm). Data were processed as described above. Guanidinium-induced denaturation measurements were performed after overnight incubation at 20 °C with increasing concentrations of guanidinium chloride (99.5% purity, Fluka), and the data were collected and processed as described above. Measurements of designed armadillo repeat proteins were performed in 20 mM Tris–HCl and 50 mM NaCl (pH 8.0). CD spectra and denaturation curves of the armadillo domain of β-catenin were measured in 50 mM Tris–HCl and 500 mM NaCl (pH 8.0). CD spectra and denaturation curves of importin-α1 were measured in 50 mM phosphate, 500 mM NaCl, and 5 mM DTT (pH 7.4). CD spectra were analyzed using CDpro.[41] Among the algorithms available in CDpro, CDSSTR was chosen for the analysis, with the reference protein set SDP48 (IBasis=7).

ANS binding

ANS fluorescence was measured using a PTI QM-2000-7 fluorimeter (Photon Technology International, USA). The measurements were performed at 20 °C in 20 mM Tris–HCl, 50 mM NaCl, and 100 μM ANS (pH 8.0) using purified proteins at a final concentration of 10 μM. ANS binding to the armadillo domain of β-catenin was measured in 50 mM Tris–HCl, 500 mM NaCl, and 100 μM ANS (pH 8.0) to avoid possible aggregation problems. The emission spectrum from 400 to 650 nm (1 nm/s) was recorded with an excitation wavelength of 350 nm. For each sample, three spectra were recorded and averaged.

Rotamer sampling of hydrophobic core mutants

A computational approach at the atomic level of detail was used to optimize the hydrophobic core. The approach

uses cycles of energy minimization and heating by molecular dynamics to sample favorable arrangements of the buried side chains and estimate the packing efficiency of residues in the hydrophobic core of a given mutant. The number of possible mutations in each repeat is 432 (Fig. 6b). Three x-ray structures were chosen as starting models to improve sampling: importin-α from *S. cerevisiae* (PDB ID 1EE4[46]) and mouse (PDB ID 1Q1T[47]), consisting each of 8 internal repeats and 2 capping repeats, and murine β-catenin (PDB ID 2BCT[22]), which consists of 10 internal repeats and 2 capping repeats. The original capping repeats of the three structures were substituted with Ny and Ca capping repeats (Fig. 3) in the models. Each mutation was modeled by deleting the side chains at the core positions of each repeat and substituting them with the new side chains with random rotamer conformations; the resulting structure was minimized to eliminate clashes. Three models (from the three initial structures) were prepared for each of the 432 combinations of allowed mutations. All the repeats in each model were designed to have the same mutation pattern.

The extended atom approximation (param19) of the CHARMM force field[71] with a distance-dependent dielectric function was used for both energy minimization and heating by short molecular dynamics runs. All the side-chain atoms not directly in contact with core residue atoms (i.e., those more than 5 Å away, in the initial conformation, from any atom of the 16 core residues of each repeat) and all the backbone atoms were restrained using a harmonic potential with a force constant of 1.0 kcal·mol^{-1}·Å^{-2}. As a consequence, only the side-chain rotatable bonds of the core residues were fully flexible. The system was further minimized in the presence of the harmonic potential. A heating–quench protocol was iterated 100 times for each of the 432 mutants and each of the three protein models (Fig. 7). The first step was a 10-ps heating to 400 K, omitting the nonbonding energy terms (i.e., van der Waals and electrostatics). The second step was a minimization including all energy terms. The aim of the heating phase was to shuffle the flexible side-chain rotamers. The absence of nonbonding energy terms and the high temperature granted a more efficient exploration of the energy landscape. After the heating step, the minimization was used to reach the nearest minimum of the total potential energy. The coordinates were stored at the end of each minimization, and a total of 100 conformations were generated for each mutant. These conformations were further minimized without the aforementioned restraints, and the potential energy was evaluated for ranking.

The CHARMM potential energy is:

$$E = E_{\text{bonding}} + E_{\text{vdw}} + E_{\text{elec}} \quad (1)$$

where E_{bonding} is the sum of bond, angle, improper, and dihedral potential terms; E_{vdw} is the van der Waals energy; and E_{elec} is the coulombic energy. The E_{elec} term was neglected for ranking, because it is insensitive to aliphatic-to-aliphatic mutations in the extended atom representation. Moreover, because of the restraints, the restricted flexibility of the backbone polar groups results in a noisy coulombic energy. Therefore, a reduced potential energy was used for ranking. For each starting structure, the energy value for the conformation i of mutant m is:

$$E_i^m = E_{i,\text{bonding}}^m + E_{i,\text{vdw}}^m \quad (2)$$

As each conformation has a different potential energy, median, first percentile (of most favorable values), and minimum energies were extracted from the energy series of the 100 conformers to characterize each mutant: these values were used to make three independent ranks. At the end of this procedure, for each of the three initial structures, three rank numbers (corresponding to median, first percentile, and minimum ranking) were assigned to each mutant, and these nine rank numbers were summed. Finally, this sum was used for the overall rank of the mutant (Supplementary Table S3). The combination of multiple structures and different scoring criteria (i.e., median, first percentile, and minimum) was used to take into account, in an approximate way, the limited sampling. The central processing unit time required for each starting model of a mutant was approximately 5 h for importin structures and 7 h for catenin structures on a single processor of a 2800-MHz Opteron dual core. The total calculation time of approximately 8000 h was distributed over 150 central processing units.

NMR

Proteins for NMR studies were produced using *E. coli* strain M15 (Qiagen) containing the plasmid pREP4 growing in minimal medium with ^{15}N-labeled ammonium chloride as the only nitrogen source. The medium was supplemented with trace metals, 150 μM thiamin, and 30 μg/ml kanamycin. Expression and purification by IMAC and gel filtration were performed as described. The buffers used for NMR measurements contained 20 mM deuterated Tris–HCl and 30 mM NaCl (pH values of 6, 7, 8, 9, 10, or 11). YC_4A and YM_4A were concentrated to 0.6 mM for NMR measurements.

Proton–nitrogen correlation maps were derived from [^{15}N,^{1}H]-HSQC experiments[72] utilizing pulsed-field gradients for coherence selection and quadrature detection[73] and incorporating the sensitivity enhancement element of Rance and Palmer.[51,74] The ^{15}N{^{1}H}-NOE data were measured using a proton-detected version of the ^{15}N{^{1}H} steady-state heteronuclear Overhauser effect.[75] All experiments were recorded on a Bruker AV 700-MHz spectrometer equipped with a triple-resonance cryoprobe at 310 K. Spectra were processed and analyzed in the spectrometer software TOPSPIN 1.3 and calibrated relative to the water resonance at 4.63 ppm proton frequency, from which the ^{15}N scale was calculated indirectly.

ELISA

Biotinylated pD-peptide fusion proteins were immobilized on NeutrAvidin-coated plates after IMAC purification using 200 μl of 10-μM protein solutions and 1 h incubation time. One hundred microliters of 1 μM armadillo repeat proteins was incubated for 1 h. Binding was detected with an anti-$MRGSH_4$ antibody (Qiagen), a secondary anti-mouse immunoglobulin G alkaline phosphatase conjugate (Sigma), and *p*-nitrophenylphosphate (Fluka). Absorbance at 405 nm was measured using a Perkin Elmer HTS 7000 Plus plate reader. A buffer solution containing 50 mM Tris–HCl, 150 mM NaCl, and 0.5% bovine serum albumin (pH 7.4) was used for all the proteins and for the blocking steps. Washing after each step was carried out with $TBST_{150}$ (Tris–HCl 50 mM, NaCl 150 mM, and 0.05% Tween 20, pH 7.4). All steps were carried out at 4 °C. Development with 4-nitrophenylphosphate and readout were performed at room temperature.

Acknowledgements

The authors want to thank W.I. Weis, M. Köhler, and E. Conti for kindly providing the plasmids containing the natural armadillo repeat protein genes. We thank Dr. P. Kolb for valuable suggestions, Dr. A. Honegger for EXCEL macros, and the other members of the Plückthun laboratory for fruitful discussions. The calculations were performed on Matterhorn, a Beowulf Linux cluster at the Informatikdienste of the University of Zürich. We thank C. Bolliger, Dr. T. Steenbock, and Dr. A. Godknecht for installing and maintaining the Linux cluster. F. Parmeggiani was the recipient of a predoctoral fellowship from the Roche Research Foundation. F.P. and G.V. are members of the Molecular Life Science Ph.D. program. This work was supported by the Swiss National Center of Competence in Research (NCCR) in Structural Biology and in part by a Discovery grant from the Kommission für Technologie und Innovation (KTI).

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2007.12.014

References

1. Hoogenboom, H. R. (2005). Selecting and screening recombinant antibody libraries. *Nat. Biotechnol.* **23**, 1105–1116.
2. Binz, H. K., Amstutz, P. & Plückthun, A. (2005). Engineering novel binding proteins from nonimmunoglobulin domains. *Nat. Biotechnol.* **23**, 1257–1268.
3. Almagro, J. C. (2004). Identification of differences in the specificity-determining residues of antibodies that recognize antigens of different size: implications for the rational design of antibody repertoires. *J. Mol. Recognit.* **17**, 132–143.
4. MacCallum, R. M., Martin, A. C. & Thornton, J. M. (1996). Antibody–antigen interactions: contact analysis and binding site topography. *J. Mol. Biol.* **262**, 732–745.
5. Marchalonis, J. J., Adelman, M. K., Robey, I. F., Schluter, S. F. & Edmundson, A. B. (2001). Exquisite specificity and peptide epitope recognition promiscuity, properties shared by antibodies from sharks to humans. *J. Mol. Recognit.* **14**, 110–121.
6. Wilson, I. A., Ghiara, J. B. & Stanfield, R. L. (1994). Structure of anti-peptide antibody complexes. *Res. Immunol.* **145**, 73–78.
7. Kuriyan, J. & Cowburn, D. (1997). Modular peptide recognition domains in eukaryotic signaling. *Annu. Rev. Biophys. Biomol. Struct.* **26**, 259–288.
8. Esteban, O. & Zhao, H. (2004). Directed evolution of soluble single-chain human class II MHC molecules. *J. Mol. Biol.* **340**, 81–95.
9. Blatch, G. L. & Lässle, M. (1999). The tetratricopeptide repeat: a structural motif mediating protein–protein interactions. *BioEssays*, **21**, 932–939.
10. Coates, J. C. (2003). Armadillo repeat proteins: beyond the animal kingdom. *Trends Cell Biol.* **13**, 463–471.
11. Smith, T. F., Gaitatzes, C., Saxena, K. & Neer, E. J. (1999). The WD repeat: a common architecture for diverse functions. *Trends Biochem. Sci.* **24**, 181–185.
12. Peifer, M., Berg, S. & Reynolds, A. B. (1994). A repeating amino acid motif shared by proteins with diverse cellular roles. *Cell*, **76**, 789–791.
13. Hatzfeld, M. (1999). The armadillo family of structural proteins. *Int. Rev. Cytol.* **186**, 179–224.
14. Harris, T. J. & Peifer, M. (2005). Decisions, decisions: beta-catenin chooses between adhesion and transcription. *Trends Cell Biol.* **15**, 234–237.
15. Anastasiadis, P. Z. & Reynolds, A. B. (2000). The p120 catenin family: complex roles in adhesion, signaling and cancer. *J. Cell Sci.* **113**, 1319–1334.
16. Nathke, I. S. (2004). The adenomatous polyposis coli protein: the Achilles heel of the gut epithelium. *Annu. Rev. Cell Dev. Biol.* **20**, 337–366.
17. Goldfarb, D. S., Corbett, A. H., Mason, D. A., Harreman, M. T. & Adam, S. A. (2004). Importin alpha: a multipurpose nuclear-transport receptor. *Trends Cell Biol.* **14**, 505–514.
18. Wieschaus, E., Nüsslein-Volhard, C. & Jürgens, G. (1984). Mutations affecting the pattern of the larval cuticle in *Drosophila melanogaster*. 3. Zygotic loci on the X-chromosome and 4th chromosome. *Wilhelm Roux's Arch. Dev. Biol.* **193**, 296–307.
19. Riggleman, B., Wieschaus, E. & Schedl, P. (1989). Molecular analysis of the armadillo locus: uniformly distributed transcripts and a protein with novel internal repeats are associated with a *Drosophila* segment polarity gene. *Genes Dev.* **3**, 96–113.
20. Groves, M. R. & Barford, D. (1999). Topological characteristics of helical repeat proteins. *Curr. Opin. Struct. Biol.* **9**, 383–389.
21. Kobe, B. & Kajava, A. V. (2000). When protein folding is simplified to protein coiling: the continuum of solenoid protein structures. *Trends Biochem. Sci.* **25**, 509–515.
22. Huber, A. H., Nelson, W. J. & Weis, W. I. (1997). Three-dimensional structure of the armadillo repeat region of beta-catenin. *Cell*, **90**, 871–882.
23. Conti, E., Uy, M., Leighton, L., Blobel, G. & Kuriyan, J. (1998). Crystallographic analysis of the recognition of a nuclear localization signal by the nuclear import factor karyopherin alpha. *Cell*, **94**, 193–204.
24. Catimel, B., Teh, T., Fontes, M. R., Jennings, I. G., Jans, D. A., Howlett, G. J. *et al.* (2001). Biophysical characterization of interactions involving importin-alpha during nuclear import. *J. Biol. Chem.* **276**, 34189–34198.
25. Forrer, P., Stumpp, M. T., Binz, H. K. & Plückthun, A. (2003). A novel strategy to design binding molecules harnessing the modular nature of repeat proteins. *FEBS Lett.* **539**, 2–6.
26. Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R. D. & Bairoch, A. (2003). ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res.* **31**, 3784–3788.
27. Binz, H. K., Stumpp, M. T., Forrer, P., Amstutz, P. & Plückthun, A. (2003). Designing repeat proteins: well-expressed, soluble and stable proteins from combinatorial libraries of consensus ankyrin repeat proteins. *J. Mol. Biol.* **332**, 489–503.
28. Mosavi, L. K., Minor, D. L., Jr & Peng, Z. Y. (2002). Consensus-derived structural determinants of the ankyrin repeat motif. *Proc. Natl Acad. Sci. USA*, **99**, 16029–16034.
29. Main, E. R., Xiong, Y., Cocco, M. J., D'Andrea, L. & Regan, L. (2003). Design of stable alpha-helical arrays from an idealized TPR motif. *Structure*, **11**, 497–508.

30. Stumpp, M. T., Forrer, P., Binz, H. K. & Plückthun, A. (2003). Designing repeat proteins: modular leucine-rich repeat protein libraries based on the mammalian ribonuclease inhibitor family. *J. Mol. Biol.* **332**, 471–487.
31. Interlandi, G., Wetzel, S. K., Settanni, G., Plückthun, A. & Caflisch, A. (2008). Characterization and further stabilization of designed ankyrin repeat proteins by combining molecular dynamics simulations and experiments. *J. Mol. Biol.* **375**, 837–854.
32. Andrade, M. A., Petosa, C., O'Donoghue, S. I., Müller, C. W. & Bork, P. (2001). Comparison of ARM and HEAT protein repeats. *J. Mol. Biol.* **309**, 1–18.
33. Schultz, J., Milpetz, F., Bork, P. & Ponting, C. P. (1998). SMART, a simple modular architecture research tool: identification of signaling domains. *Proc. Natl Acad. Sci. USA*, **95**, 5857–5864.
34. Letunic, I., Copley, R. R., Pils, B., Pinkert, S., Schultz, J. & Bork, P. (2006). SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res.* **34**, D257–D260.
35. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680.
36. Lange, A., Mills, R. E., Lange, C. J., Stewart, M., Devine, S. E. & Corbett, A. H. (2007). Classical nuclear localization signals: definition, function, and interaction with importin alpha. *J. Biol. Chem.* **282**, 5101–5105.
37. Xu, W. & Kimelman, D. (2007). Mechanistic insights from structural studies of beta-catenin and its binding partners. *J. Cell Sci.* **120**, 3337–3344.
38. von Kries, J. P., Winbeck, G., Asbrand, C., Schwarz-Romond, T., Sochnikova, N., Dell'Oro, A. *et al.* (2000). Hot spots in beta-catenin for interactions with LEF-1, conductin and APC. *Nat. Struct. Biol.* **7**, 800–807.
39. Hoffmans, R. & Basler, K. (2004). Identification and *in vivo* role of the Armadillo–Legless interaction. *Development*, **131**, 4393–4400.
40. Leung, S. W., Harreman, M. T., Hodel, M. R., Hodel, A. E. & Corbett, A. H. (2003). Dissection of the karyopherin alpha nuclear localization signal (NLS)-binding groove: functional requirements for NLS binding. *J. Biol. Chem.* **278**, 41947–41953.
41. Sreerama, N. & Woody, R. W. (2004). Computation and analysis of protein circular dichroism spectra. *Methods Enzymol.* **383**, 318–351.
42. Slavik, J. (1982). Anilinonaphthalene sulfonate as a probe of membrane composition and function. *Biochim. Biophys. Acta*, **694**, 1–25.
43. Ptitsyn, O. B. (1995). Molten globule and protein folding. *Adv. Protein Chem.* **47**, 83–229.
44. Butterfoss, G. L. & Kuhlman, B. (2006). Computer-based design of novel protein structures. *Annu. Rev. Biophys. Biomol. Struct.* **35**, 49–65.
45. Voigt, C. A., Gordon, D. B. & Mayo, S. L. (2000). Trading accuracy for speed: a quantitative comparison of search algorithms in protein sequence design. *J. Mol. Biol.* **299**, 789–803.
46. Conti, E. & Kuriyan, J. (2000). Crystallographic analysis of the specific yet versatile recognition of distinct nuclear localization signals by karyopherin alpha. *Structure*, **8**, 329–338.
47. Fontes, M. R., Teh, T., Toth, G., John, A., Pavo, I., Jans, D. A. & Kobe, B. (2003). Role of flanking sequences and phosphorylation in the recognition of the simian-virus-40 large T-antigen nuclear localization sequences by importin-alpha. *Biochem. J.* **375**, 339–349.
48. Chothia, C. (1975). Structural invariants in protein folding. *Nature*, **254**, 304–308.
49. Baum, J., Dobson, C. M., Evans, P. A. & Hanley, C. (1989). Characterization of a partly folded protein by NMR methods—studies on the molten globule state of guinea-pig alpha-lactalbumin. *Biochemistry*, **28**, 7–13.
50. Dyson, H. J. & Wright, P. E. (1998). Equilibrium NMR studies of unfolded and partially folded proteins. *Nat. Struct. Biol.* **5**, 499–503.
51. Palmer, A. G. (2001). NMR probes of molecular dynamics: overview and comparison with other techniques. *Annu. Rev. Biophys. Biomol. Struct.* **30**, 129–155.
52. Kalderon, D., Roberts, B. L., Richardson, W. D. & Smith, A. E. (1984). A short amino acid sequence able to specify nuclear location. *Cell*, **39**, 499–509.
53. Choi, H. J. & Weis, W. I. (2005). Structure of the armadillo repeat domain of plakophilin 1. *J. Mol. Biol.* **346**, 367–376.
54. Mosavi, L. K. & Peng, Z. Y. (2003). Structure-based substitutions for increased solubility of a designed protein. *Protein Eng.* **16**, 739–745.
55. Main, E. R., Stott, K., Jackson, S. E. & Regan, L. (2005). Local and long-range stability in tandemly arrayed tetratricopeptide repeats. *Proc. Natl Acad. Sci. USA*, **102**, 5721–5726.
56. Kajander, T., Cortajarena, A. L., Main, E. R., Mochrie, S. G. & Regan, L. (2005). A new folding paradigm for repeat proteins. *J. Am. Chem. Soc.* **127**, 10188–10190.
57. Wetzel, S. K., Settanni, G., Kenig, M., Binz, H. K. & Plückthun, A. (2008). Folding and unfolding mechanism of highly stable full consensus ankyrin repeat proteins. *J. Mol. Biol.* In press. doi:10.1016/j.jmb.2007.11.046
58. Eriksson, A. E., Baase, W. A., Zhang, X. J., Heinz, D. W., Blaber, M., Baldwin, E. P. & Matthews, B. W. (1992). Response of a protein structure to cavity-creating mutations and its relation to the hydrophobic effect. *Science*, **255**, 178–183.
59. Schueler-Furman, O., Wang, C., Bradley, P., Misura, K. & Baker, D. (2005). Progress in modeling of protein structures and interactions. *Science*, **310**, 638–642.
60. Desjarlais, J. R. & Handel, T. M. (1999). Side-chain and backbone flexibility in protein core design. *J. Mol. Biol.* **290**, 305–318.
61. Looger, L. L., Dwyer, M. A., Smith, J. J. & Hellinga, H. W. (2003). Computational design of receptor and sensor proteins with novel functions. *Nature*, **423**, 185–190.
62. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H. *et al.* (2000). The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242.
63. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.
64. McGinnis, S. & Madden, T. L. (2004). BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.* **32**, W20–W25.
65. Guex, N. & Peitsch, M. C. (1997). SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis*, **18**, 2714–2723.
66. Koradi, R., Billeter, M. & Wüthrich, K. (1996). MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graphics*, **14**, 51–55, 29–32.
67. Sambrook, J. & Russell, D. W. (2001). *Molecular Cloning: A Laboratory Manual*, 3rd edit., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
68. Inoue, H., Nojima, H. & Okayama, H. (1990). High efficiency transformation of *Escherichia coli* with plasmids. *Gene*, **96**, 23–28.

69. Köhler, M., Speck, C., Christiansen, M., Bischoff, F. R., Prehn, S., Haller, H. *et al.* (1999). Evidence for distinct substrate specificities of importin alpha family members in nuclear protein import. *Mol. Cell. Biol.* **19**, 7782–7791.
70. Cull, M. G. & Schatz, P. J. (2000). Biotinylation of proteins *in vivo* and *in vitro* using small peptide tags. *Methods Enzymol.* **326**, 430–440.
71. Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. & Karplus, M. (1983). CHARMM—a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **4**, 187–217.
72. Bodenhausen, G. & Ruben, D. J. (1980). Natural abundance N-15 NMR by enhanced heteronuclear spectroscopy. *Chem. Phys. Lett.* **69**, 185–189.
73. Keeler, J., Clowes, R. T., Davis, A. L. & Laue, E. D. (1994). Pulsed-field gradients: theory and practice. *Methods Enzymol.* **239**, 145–207.
74. Kay, L. E., Keifer, P. & Saarien, T. (1992). Pure absorption gradient enhanced heteronuclear single-quantum correlation spectroscopy with improved sensitivity. *J. Am. Chem. Soc.* **114**, 10663–10665.
75. Noggle, J. H. & Schirmer, R. E. (1971). *The Nuclear Overhauser Effect: Chemical Applications*, Academic Press, New York.

8 A fast implicit solvent model for proteins and lipids

A class of membrane-associated peptides are observed to assume regular secondary structure upon binding to lipid membranes. This process is the basis for all mechanisms of action of toxins and microbial peptides, and it is essential for the stability of membrane proteins [89]. Furthermore understanding the interactions that determine peptide orientation and stability within membrane might help the engineering of membrane protein [90]. Partitioning and folding of a membrane-associated peptides may be represented by a thermodynamic cycle, which generalizes the two state folding equilibrium expressed by the kinetic scheme (1):

$$\begin{array}{ccc} U & \longleftrightarrow & F \\ \updownarrow & & \updownarrow \\ U^* & \longleftrightarrow & F^* \end{array}$$

where states F and F^* are the membrane unbound and membrane bound folded forms respectively, and the states U and U^* are the unfolded ones. The states F and U^* are marginally populated, therefore only the equilibrium $U \longleftrightarrow F^*$ is observed in experiments [91]. Nevertheless, structural characterization of the intermediate states and the interaction that drives the membrane association and folding are fundamental for the reasons mentioned above.

The aim of the work exposed in section 8.1 is to simulate the association and the folding of the amphipathic melittin peptide in presence of detergent micelles, in analogy with experiments. To achieve this task a new implicit solvent , which overcome some limitations of the models currently used, has been developed. The solvation model described by eq. (6) has the drawback that it doesn't include formal charges. Furthermore surface tension parameters σ_i don't have a precise physical meaning, which leads to parameterization difficulties. For these reasons a novel model based on solvent accessible surface has been developed. Here the polar and the non polar surfaces contributions are uncoupled, and the polar term is derived from first shell approximation of solvation energy of a charge q, as described in section 8.1:

$$G_{solv} = \sigma_{np} \sum_i S_i + \sigma_p \sum_i \left(\frac{q_i^2}{R_i + r_p} \frac{S_i}{S_i^{free}} \right)$$

The advantage of this form for the solvation energy is that the parameters acquires a physical meaning: σ_{np} is the non-polar surface tension of the molecule, and σ_p is the polar surface tension.

Using this implicit model of aqueous solvent, molecular dynamics simulations of melittin in presence of explicit dodecylphosphocholine were performed to investigate how the micelle environment influences melittin structure. Both constant temperature and replica exchange approaches were adopted (see section 3.2.2). In the former simulations, reversible transitions from unstructured conformations to helical fold, when melittin is bound to the micelle, were reproduced. In the latter it was found that a specific number of lipids are needed to stabilize the helical fold. The simulation results confirmed the proposed structural model of melittin-micelle complex, supporting a solid microscopic view of helical formation.

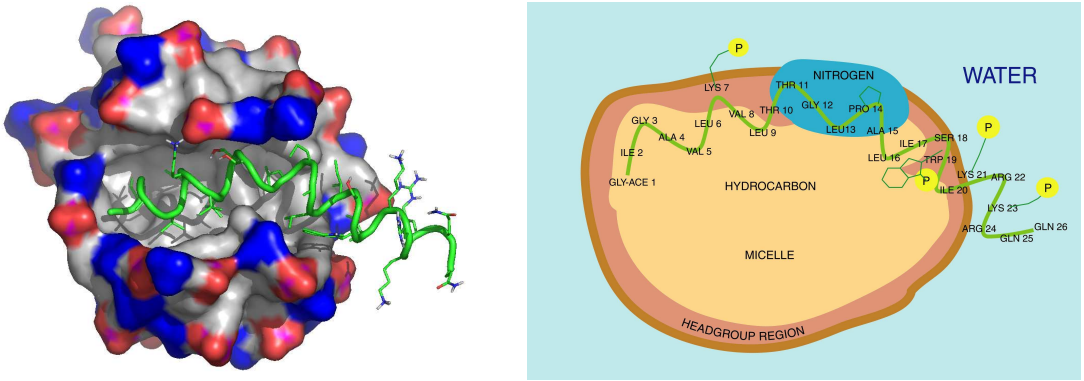


Figure 6: Left: representative melittin structure in DPC micelle obtained by simulation. Right: schematic of the amino acids partitioning into the micelle environments: headgroup phosphates coordinating with melittin sidechains are represented by a yellow circle, choline nitrogens coordinating with backbone in the kink region are represented in blue.

8.1 Folding of Helical Peptide at the Micelle-Water Interface. [Manuscript in preparation].

Folding of helical peptide at the micelle-water interface

Riccardo Pellarin and Amedeo Caflisch

Department of Biochemistry
University of Zürich
Winterthurerstrasse 190
CH-8057 Zürich, Switzerland
Phone: (+41 44) 635 55 21
FAX: (+41 44) 635 68 62
email: pellarin@bioc.unizh.ch

July 4, 2007

Abstract

Lipid micelles, as well as membranes, are known to strongly promote secondary structure formation in a wide range of membrane active peptides. This process is the basis for a variety of fundamental biological functions, including signal transduction, toxicity and folding of membrane protein. However, little is known on the detailed mechanisms and the driving forces underlying this process. Among many different peptides which associate to membranes, melittin is the most investigated one. It is largely unstructured when free in solution, but clearly adopts an amphipathic α-helical conformation when partitioned into membranes [1]. In this work molecular dynamics simulations are used to investigate the interaction and the folding of melittin at the dodecyl-phosphocholine (DPC) micelle surface. An implicit solvent model with explicit lipid molecules and peptide, allows the microsecond time scale sampling, capturing reversible folding events of the helical peptide. Both constant temperature and replica exchange simulations results mostly agrees with present structural models of melittin-membrane interactions, supporting a solid microscopic view of helix formation on membrane surface.

1 Introduction

Membrane associated peptides form a wide class of molecules that includes sequences known to be toxic or functionally relevant for the cell. Furthermore this class of peptides has been adopted as a model for large membrane proteins, and their investigation contributes to the understanding of the mechanisms of membrane-proteins interactions. Melittin, the major component of the honey bee *Apis mellifera* venom, is one of the best known member of this class. It is a 26 amino acids peptide which has a potent hemolytic activity and induces membrane leakage [2, 3]. Melittin is strongly amphipathic, a feature that characterizes many membrane active peptides such as hormones [4, 5], antibiotics [6, 7] or designed sequences [8]. Its sequence is:

$$+\mathbf{G}\text{IGAVL}\mathbf{K}\text{VLT}_{10}\text{TGLPALISWI}_{20}\mathbf{KRKR}\text{QQ-NH2}$$

where the bold residue letters are the charged residues. It is a strongly basic sequence with six positive charges; the N-terminal amino group, Lys at position 7 and four charges in the highly basic C-terminal segment.

Structural investigations of this peptide have been performed in several conditions. Melittin in aqueous solution is monomeric with a ^{1}H-NMR spectrum close to random coil, though fragments of the polypeptide chain might adopt non-random spatial strucure [9, 10].

The high resolution structure of melittin has been resolved by X-ray crystallography [11, 12] starting from aqueous solution under strong ionic conditions, where melittin adopts a homotetramer oligomeric form. In the crystal melittin is mostly α-helical (see figure 1 left) with just the exception of proline 14, which missing the amide that ideally interacts with Thr-10 carbonyl, breaks the α-helical structure. Hence the helix is kinked, and the axes of the two sub-helices, defined by residues 1-10 and 16-26, intersect with an angle of about 120 degrees.

Structures of monomeric melittin in methanol and in dodecylphosphocholine micelles have been determined using proton NMR and amide exchange analysis.

In methanol melittin is monomeric and α-helical [3]. The kink angle in this case, of 160 degrees, is considerably larger than the value found for the crystal [13]. In presence of perdeuterated phosphatidylcholine micelles melittin is still α-helical, with an estimated kink of 135 degrees [14]. Amides on the hydrophobic side of the amphipathic helix show the largest chemical shift changes compared with the amides of melittin in methanol, and the slowest rates, supporting a surface location for the peptide with the non-polar side of the helix facing the micelle interior. In magnetically oriented DMPC bilayer, the conformation orientation and dynamics of melittin have been determinad using NMR spectroscopy [15], where the peptide was found to adopt a transmembrane α-helical conformation, with a kink angle of 140-160 degrees. In fluorinated alcohols melittin shows a kink angle of 73 degrees [16], while in hexafluoroacetone the structure is closest to the crystal and micellar structure, displaying a kink of 144 degrees [17].

Pro-14, in the hinge region between the two helical segments, is believed to have a particular structural role [18]. The proline ring produces steric hindrance to the straight α-helical conformation as a result of not having a NH group available for a hydrogen bond. In fact substitution of L-Pro with its diasteroisomer D-Pro contributes to increase the misaaligniament of the hydrophobic faces of the helix, and sensibly reduced the hemolitic activity. NMR structural analysis [19] provided structural models for D-melittin.

The molecular mechanism of melittin-induced cell lysis is controversial, due to instrinsic difficulties in characterizing the interactions between the peptide and the membrane. Thanks to the atomistic details of molecular dynamics, the knowledge in this field has been broaden. Explicit solvent molecular dynamics of membrane proteins [20–23] and peptides [24,25] has the great advantage of having an accurate force field. A number of fully atomistic molecular dynamics simulations of melittin have been performed in different solvents [26], in bilayers [26–28] or in pore forming oligomers [29] and evidentiate conformational changes and orientation in membrane of the peptide. Some investigators have studied the configuration and

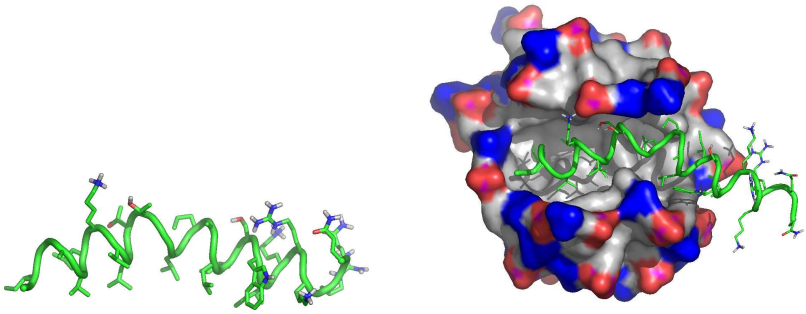


Figure 1: Left: X-ray structure of melittin in tetrameric form [12]. Right: representative melittin structure in DPC micelle from simulations. DPC molecules are diplayed as a surface, the carbons of the tail are colored in gray, while the choline and the phosphate are blue and red respectively.

the stability of a single membrane pore bound by four melittin molecules in a fully hydrated membrane (Jung-Hsin and Baumgartner Biophys Journal vol 78 2000 1714). Fully atomistic simulations revealed the complexity behind polypeptide chain and specific lipid components interactions. Yet, explicit solvent simulation is limited in conformational sampling, especially for slow events such as peptide binding and folding. Many theoretical works on membrane proteins propose the implicit treatment of a membrane bilayer [30–33]. The membrane is modelled as a region with low dielectric continuum, and the interaction with the solute are included into the force field using a mean field approximation. These models are very efficient in terms of calculations, but might be innaccurate, underestimating relevant interactions between proteins and lipids. Membrane insertion of peptides has been intesively investigated using molecular mechanics methods.

The present study was motivated by two main questions: Is it possible to simulate the spontaneous folding of melittin into a micelle environment? Can we reproduce the structural models of melittin? We developed an efficent implicit water model with explicit lipids that has an intermediate accuracy between the explicit water and the totally implicit models. This solvation model, explained in the methods part, extends the SASA models largely adopted for small peptides folding, by including formal charges and lipid molecules. The molecular dynamics

simulations were in the microsecond timescale, an essential time lapse to capture the slow transitions that lead to the helical conformation. Results from constant temperature and replica exchange simulations are in good agreement with the experimental observations.

2 Results

Constant temperature simulations. Simulations at a temperature of 330 K have been performed with and without lipids. The simulation results in the absence of lipids are needed to compare the structural changes induced by melittin-lipid interactions.

Melittin in pure solvent. At 330 K, in pure solvent, melittin has a tendency to form helical structures, but the central region (around the proline 14) is rather unstructured, confering flexibility to the whole peptide. During the 1 μs simulation the sampled conformations are distant from the X-ray structure, as confirmed by rmsd analysis displayed in figure 2A. The average RMSD value is 6 Å, and the kink angle, measured as the instantenous angle between the axis of the helices formed in the segment 1-14 and 15-26, highly fluctuates in the range 0-180. The helical contacts analysis reveals a small propensity to form hydrogen bonds in the proline region. Helical contacts between residues 11 and 15, and 12 and 16 are never formed during the simulation. Under these conditions, melittin does not assume any preferential structure, as shown by conformational clustering analysis using fingerprints of contacts and kink angle, as described in the methods part. The cluster with the highest number of visits is just marginally populated, with only 4% of the saved snapshots. The sampled conformations have a high helical kink, mainly produced by hydrophobic collapse.

Folding of melittin in micelle. Three 0.5 μs long simulations were started froma fully extended conformation of melittin surrounded by 40 monodispersed DPC molecules at the temperature of 330 K. The concentration of DPC is $4{\cdot}10^{-5}$ DPC/Å^3. In the early steps of the simulation the lipids aggregate, and within 10

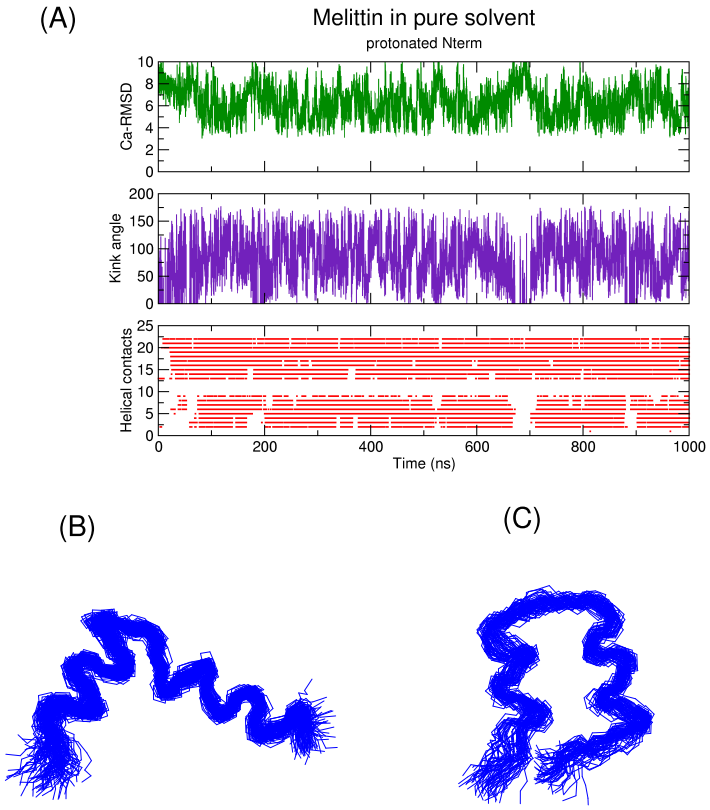


Figure 2: Melittin in pure solvent: (A) Time series of the C_α-RMSD with respect to the X-ray structure of figure 1 left. The kink angle is measured as described in the methods part. The helical contacts between residue i and i+4, described in the methods part, are represented by a red dot. (B) Most populated cluster in pure solvent (4%). (C) second most populated cluster in pure solvent (1%).

ns a spherical micelle is spontaneously produced. As described in the methods part, the micelle structure consists of an hydrophobic core, that hosts the lipid tails, and a hydrophilic surface, consisting of phosphate and choline moieties (see figure 12). Within the first 50 ns the peptide reaches the micelle, and binds onto the micelle surface. The time series of the C_α-RMSD with respect to the X-ray structure, helical contacts and the kink angle is showed in figure 3A. In the time intervals emphasized in gray in figure 3A, the C_α-RMSD with respect to the crystal structure is less than 4.0 Å , and the kink angle has a well defined value of 140 ± 20 degrees, a range that agrees with the 135 ± 15 found in micelle using 2D-NMR

and distance geometry calculations [14]. The helical contacts analysis displayed in figure 3 confirms that the peptide has a full helical conformation, which extends also in the Pro-14 region. On the contrary to what happens in pure solvent, in presence of lipids the contacts between residues 11 and 15 and residues 12 and 16 are formed. The most populated cluster, 25 %, represented in figure 3B and 3C, is a collection of helical structures with different kink angles.

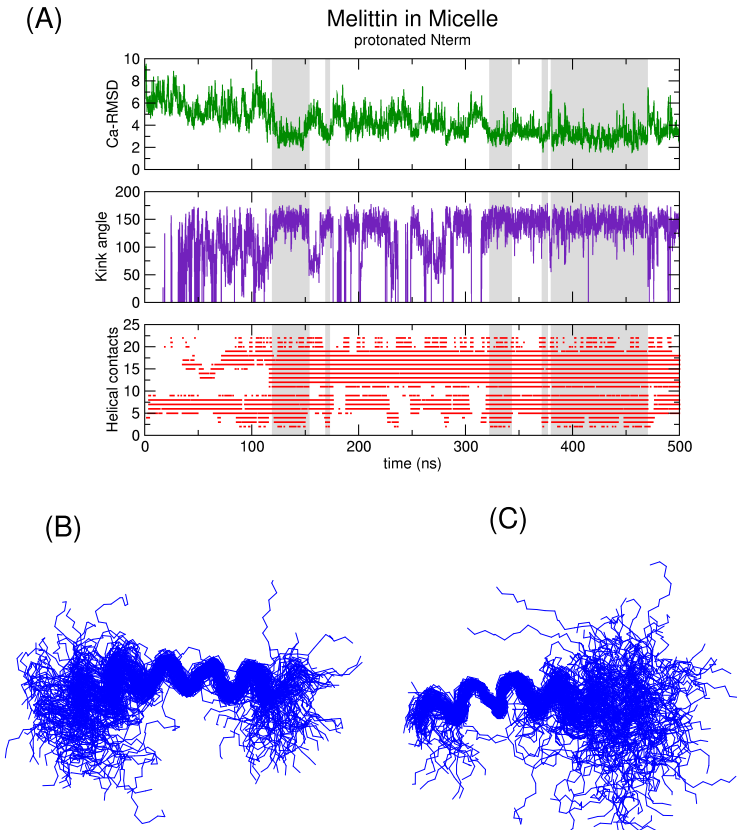


Figure 3: Melittin in DPC micelle: (A) Time series of the Cα rmsd with respect to the X-ray structure, the kink angle value and the helical contacts. (B) Most populated cluster (25 %), fitted at the C-terminal. (C) Most populated cluster, fitted at the N-terminal.

Melittin in micelle: interactions with lipid moieties. The structures belonging to the most populated cluster were analized in detail. In figure 4 the histogram of melittin-micelle atomic interactions is reported. The solvent explosure

is distributed inhomogeneously along the sequence. Segment from 21 to 26, and the N terminal segment, are more exposed to the aqueous solvent than the hydrophobic region 8-20. Among the residues belonging to this segment Val-8, Leu-9, Leu-13, Leu-16, Ile-17, Trp-19 and Ile-20 are deeply inserted in the hydrophobic core of the micelle, reflecting an amphipathic distribution of sidechains around the helix. Trp-19 displays a very high propensity to interacts with the different moieties of the DPC molecules, and in particular with the phosphate groups and the carbon groups. Therefore, Trp-19 is partitioned at the surface-bulk phase of the micelle, a motif common to many integral membrane proteins, and confirmed by NMR measurements of Trp-analogues interacting with phosphatidylcholine membranes [34]. Lysines interact preferentially with phosphates, more than arginines. In figure 5 the cylindrical projection of selected atomic distributions around the helix surface is reported. Here is evident that phosphates coordinates close to lysine sidechains, while choline preferentially coordinates with peptide moieties at the Pro-14 region.

Melittin in micelle: peptide orientation. Amide exchange analysis [35] of melittin in fully hydrated phospholipid bilayers reveals that the helix is oriented with the hydrophobic face directed toward the interior of the membrane. More recently vibrational spectroscopy studies [36] evidenciated that melittin might have both transbilayer and parallel orientation, with a preference for the parallel one. In DPC micelle, using spin labeling and ^{1}H NMR [37], melittin is located at the surface.

Lysines are exposed to the solvent and interact with phosphate groups, while the hydrophobic sidechains contact with the lipid carbon tails. N terminal amino group is completely exposed. These results enforce the view that folded melittin is located at the surface of the micelle, i.e., micelle/water interface (as depicted in figure 1 right).

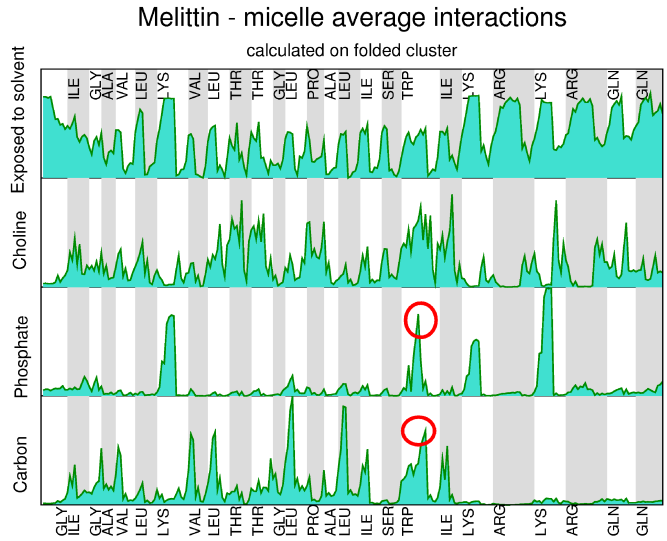


Figure 4: Histograms of melittin-micelle interactions calculated on the most populated cluster of structures. The X-axis is the atom number. Residues are labelled from the N terminal to the C terminal.

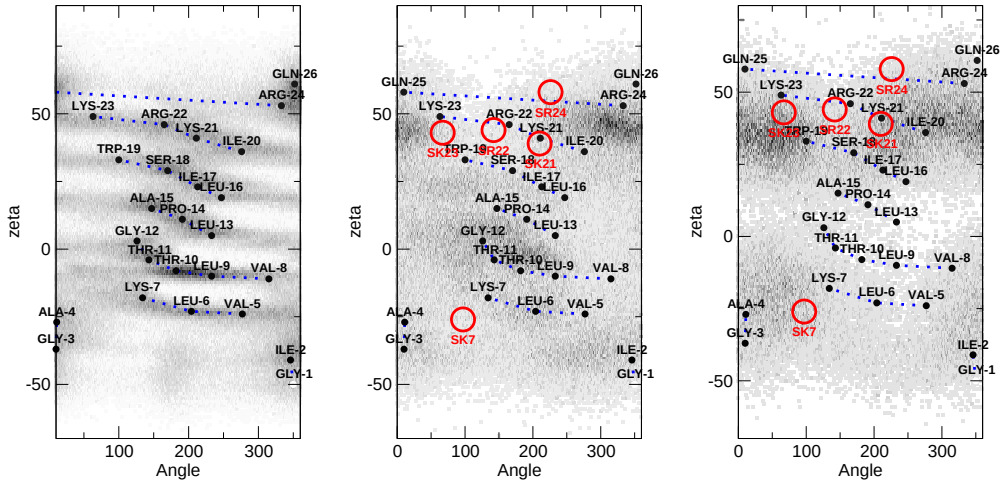


Figure 5: Cylindrical projection of the atomic distribution probabilities around the folded helical conformation. Y-axis is the position along the helical axis, X-axis is the angular position. The blue dotted lines are drawn to help the eye. The red circles represent the average positions of charged moities of the Arg and Lys sidechains (guanidinium and ammonium for Arg and Lys respectively). Darker colors are higher probabilities. Left: distribution of C_α atoms, residues are labeled here and in the other two projections. Center: choline distributions, in red the positions of the positively charged sidechains. Right: phosphate distributions.

Replica exchange simulations. 0.5 μs replica exchange simulation at three different melittin/DPC concentration ratios 1/30, 1/40 and 1/50 were performed, keeping the same lipid concentration of $4 \cdot 10^{-5}$ DPC/Å^3 for all simulations. Fourteen replicas were used (290 K to 390 K). The α-helical content increases with the size of the micelle to which is bound (figure 8). The native content (α-Helix+Kink) has a maximum at 30-40 DPCs (figure 7). The unbound peptide is fairly random coil (loop+turn+coil), in agreement with the experimental observations [1]. A native-like conformation prefers a size of 30-40 lipids. There is a deep minimum at the folded conformation only with a bound micelle of 30-40 lipids. This is consistent with NMR measurements [38].

The micelle environment influences the helical fold and viceversa. Stuctural features of micelle-bound melittin have been investigated using ^{1}H NMR, ultracentrifugation, and circular dichroism [38]. It was reported that stoichiometry of the melittin-DPC complex is about 1 peptide and 32±10 detergent molecules. Being the kink angle similar for crystalline state, methanolic solution and micelles, it was proposed that micelles conform to melittin structure, rather than the melittin conform to the curvature of the micelle [39].

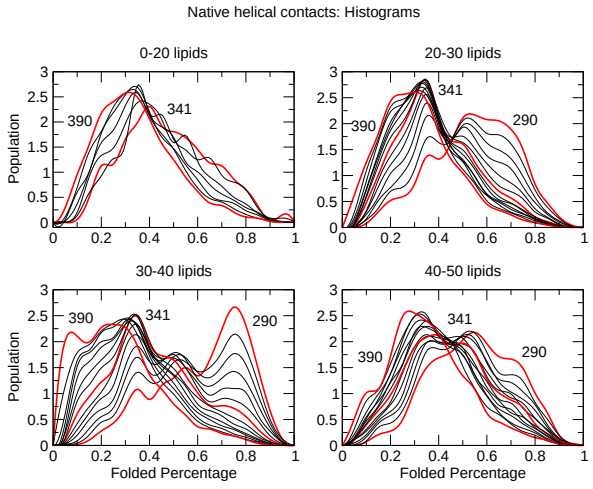


Figure 6: Histograms of the native helical contacts calculated at different micelle size and temperature. Three representative temperatures, i.e., 290, 341 and 390, are shown as red lines.

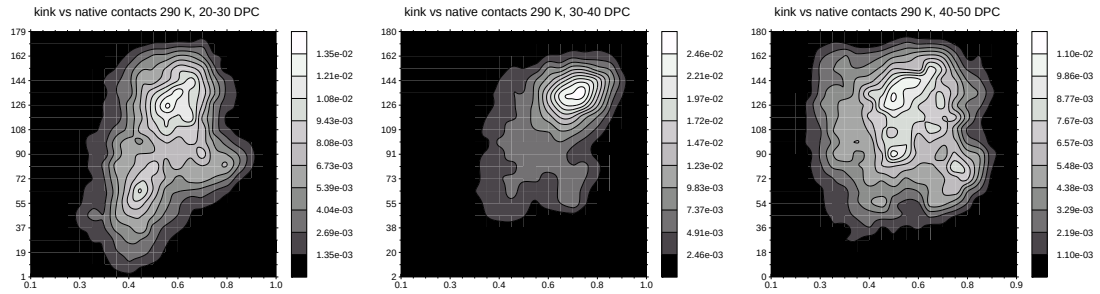


Figure 7: Contour plots of the populations of conformers with given kink angle (Y=axis) and α-helical percentage (X-axis) at three different aggregation size of the associated micelle.

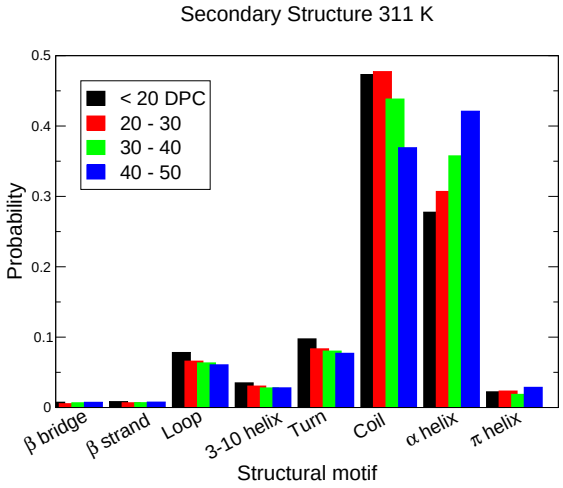


Figure 8: Secondary structure of melittin at 311 K, as a function of the number of associated lipids

3 Conclusions

Molecular dynamics simulations of the melittin peptide in presence of explicit DPC and an implicit model of aqueous solvent were carried out to investigate how micelle environment influences the structure of melittin. Constant temperature molecular dynamics reproduce a reversible transition from unstructured conformations to stable helical fold when melittin is bound to the micelle. Replica exchange simulations revealed that the the micelle aggregation number needed to stabilize the

helix conformation is between 30 to 40 lipids, a number that provides a perfect match between the hydrophobic portion of the helix and the hydrophobic core of the micelle.

The structural model of melittin that emerges from these simulations agrees with experimental data. The triptophan residue partitioning has been intesively investigated for many sequences. It is postulated that tryptophan-phosphocholine interaction may mediate important peptide conformational changes [40]. Model system consisting of phosphocholine lipids and α-helical peptides were employed to investigate the partitioning of Trp residues and Lys using a full range of biophysical experiments [41]. A functional role of Trp residues has been postulated [42], where these residues are involved in the translocation of protein through the membrane and that following translocation, Trp residues serve as anchors on the periplasmic side of the membrane.

Trp residues displace at the lipid carbonyl region, while Lys prefer to be located closer to the aqueous phase, near the lipid phosphate group. Interactions between positively charged amino-acids and phosphate group of phospholipid has been observed in a multinanosecond simulation of the N-terminal region of human pulmunary surfactant protein-B in DPPC monolayers ([43] and in Palmitic acid monolayer [44]. Spin labeling experiments at lysine positions 7, 21 and 23 assign an apparent order of accessibility to these residues, which ranks $23 < N < 21 < 7$ [45].

In this work, an efficient model was developed for use with explicit lipid molecules. It has allowed the exploration in the microsecond timescale. Explicit treatment of lipids is important for correct peptide behaviour. Possible applications are the aggregation and equilibration of lipids or surfactants on membrane proteins, and the association of TM helices.

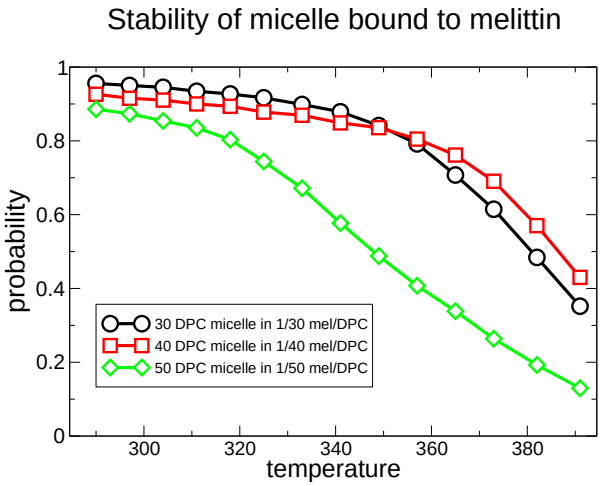


Figure 9: Temperature dependent probability that a micelle of given size is bound to mellittin in the three REM simulations.

4 Methods

Simulation protocols. All simulations presented in this work were done with the CHARMM program [46]. Constant temperature molecular dynamics have been performed using Langevin integrator with a friction of 0.15 ps^{-1} with periodic boundary conditions. Replica exchange simulations were set up according to the protocols described by Rao et al. [47]. In all simulations the starting conformation of melittin was completely extended, and lipids were monodispersed. The peptide was modeled by explicitly considering all heavy atoms and the hydrogen atoms bound to nitrogen or oxygen atoms (PARAM19 force field [46]. The CHARMM PARAM19 default cutoffs for long range interactions were used, i.e., a shift function [46] was employed with a cutoff at 7.5 Å for both the electrostatic and van der Waals terms. This cutoff length was chosen to be consistent with the parameterization of the force-field and implicit solvation model.

For both REMD and constant temperature MD, the SHAKE algorithm was used to fix the length of the covalent bonds involving hydrogen atoms, which allows an integration time step of 2 fs. Furthermore, the nonbonded interactions were updated every 10 dynamics steps and coordinate frames were saved every 20 ps for

a total of $5 \cdot 10^4$ conformations/μs. A 1 μs run requires approximately 4 weeks on a 2.0 GHz Athlon processor.

Dihedral Potential for DPC alkane chain. As an united atom force field has not been developed for lipids in CHARMM, it has been necessary to include it into the standard PARAM19. United atom dihedral potential V'_ϕ for the alkane $(CH2)_4$ chains, i.e. the dihedral potential that accounts for hydrogen presence, was derived from the all-atoms atoms CHARMM force field PARAM27 [48] using a coarse graining procedure:

$$V'_\phi(\mathbf{R}) + V'^{eff}_{vdw}(\mathbf{R}) = V_\phi(\mathbf{R}, \mathbf{r}) + V^{eff}_{vdw}(\mathbf{R}, \mathbf{r}) + V^{eff}_{elec}(\mathbf{R}, \mathbf{r})$$
$$V'^{eff}_\phi = V^{eff}_\phi$$

where $\mathbf{R}$ are the the carbons coordinates, $\mathbf{r}$ are the hydrogens coordinates, V' are the coarse grained potentials, and V are the extended atoms potentials. The effective potentials are measured thruogh:

$$V^{eff} = \frac{1}{2}\langle V_{13}\rangle + \frac{1}{2}\langle V_{24}\rangle + \langle V_{14}\rangle + \langle V_{23}\rangle$$

where 1,2,3,4 are the indicate the first, second, third and fourth $CH2$ group in the dihedral chain. The average $\langle\cdot\rangle$ is a (480K) high temperature average, performed on a simulation of a single DPC molecule simulation. Angles, bond and van der Waals parameters were assigned as standard values for the extended CH2 carbon of PARAM19. The all-atoms atoms affective potential V^{eff}_ϕ is represented in figure 12. Fourier transform of V^{eff}_ϕ yelds to the parameters for the coarse grained potential, whose first three harmonics are reported in table 1.

The dihedral populations of a simulation of the DPC molecule with the coarse grained potential $V'_\phi(\mathbf{R})$, reported in figure 10, agrees with the original potential, validating thus the whole procedure.

Solvent Model The solvation energy E_{solv} is expressed as an additive expression of the atomic solvent accessible surface (SASA):

Bond Energy

Atom types		k_b $(kcal \cdot mol^{-1} \cdot \AA^{-2})$	l_0 (Å)
NTL	CTL5	215.00	1.47 (a)
NTL	CTL4	215.00	1.47 (a)
CTL4	CTL4	222.5	1.53 (a)
CTL4	OSL	340.0	1.43 (a)
OSL	PL	270.0	1.61 (a)
O2L	PL	580	1.48 (a)
OSL	CTL2	340	1.43 (a)
CTL2	CTL2	222.5	1.53 (a)
CTL2	CTL3	222.5	1.52 (a)

Angle Energy

Atom types			k_a $(kcal \cdot mol^{-1} \cdot rad^{-2})$	θ_0 (degrees)
CTL5	NTL	CTL5	60.0	109.5 (a)
CTL5	NTL	CTL5	60.0	109.5 (a)
NTL	CTL4	CTL4	67.7	115.0 (a)
CTL4	CTL4	OSL	75.7	110.10 (a)
CTL2	CTL2	OSL	75.7	110.10 (a)
PL	OSL	CTL4	20	120 (a)
OSL	PL	O2L	98.9	111.6 (a)
OSL	PL	OSL	80	104.3 (a)
O2L	PL	O2L	120	120 (a)
CTL2	CTL2	CTL2	58.35	113.6 (a)
CTL2	CTL2	CTL3	58.35	113.6 (a)

Dihedral Energy

Atom types				k_ϕ $(kcal \cdot mol^{-1})$	n	ψ_n (degrees)
CTL5	NTL	CTL4	CTL4	0.90	3	0.00 (b)
NTL	CTL4	CTL4	OSL	1.40	3	0.00 (b)
CTL4	CTL4	OSL	PL	0.9	3	0.00 (b)
CTL2	CTL2	OSL	PL	0.9	3	0.00 (b)
CTL4	OSL	PL	OSL	0.25	3	0.00 (b)
CTL4	OSL	PL	OSL	0.75	2	0.00 (b)
CTL2	OSL	PL	OSL	0.25	3	0.00 (b)
CTL2	OSL	PL	OSL	0.75	2	0.00 (b)
OSL	CTL2	CTL2	CTL2	1.40	3	0.00 (c)
OSL	CTL2	CTL2	CTL2	0.10	2	0.00 (c)
CTL2	CTL2	CTL2	CTL2	0.52	1	0.00 (c)
CTL2	CTL2	CTL2	CTL2	1.50	3	0.00 (c)
CTL2	CTL2	CTL2	CTL2	0.13	4	0.00 (c)
CTL2	CTL2	CTL2	CTL3	0.52	1	0.00 (c)
CTL2	CTL2	CTL2	CTL3	1.50	3	0.00 (c)
CTL2	CTL2	CTL2	CTL3	0.13	4	0.00 (c)

van der Waals Energy

Atom types	E^{vdW} $(kcal \cdot mol^{-1})$	r^{vdW} $(\AA)$
CTL2	-0.1142	2.235 (d)
CTL3	-0.1811	2.165 (d)
CTL4	-0.1142	2.235 (d)
CTL5	-0.1811	2.165 (d)
O2L	-0.12	1.70 (a)
OSL	-0.1521	1.77 (a)
NTL	-0.20	1.85 (a)
PL	-0.585	2.15 (a)

Table 1: DPC force field parameters. (a) from [48], (b) from [49] (c) from dihedral reparametrization, (d) from [46].

$$
\begin{aligned}
E_{solv} &= E_{np} + E_p = \\
&= \sigma_{np} \sum_i S_i + \sigma_p \sum_i \frac{q_i^2}{R_i + r_p} \frac{S_i}{S_i^{free}}
\end{aligned} \tag{1}
$$

where i is the atom index, S_i is the atomic surface, σ_p and σ_{np} are the polar and non-polar surface tensions respectively, q_i is the partial charge of atom i, R_i is the van der Waals radius, $r_p = 1.4\mathring{A}$ is the water probe radius and S_i^{free} is the surface of the free atom:

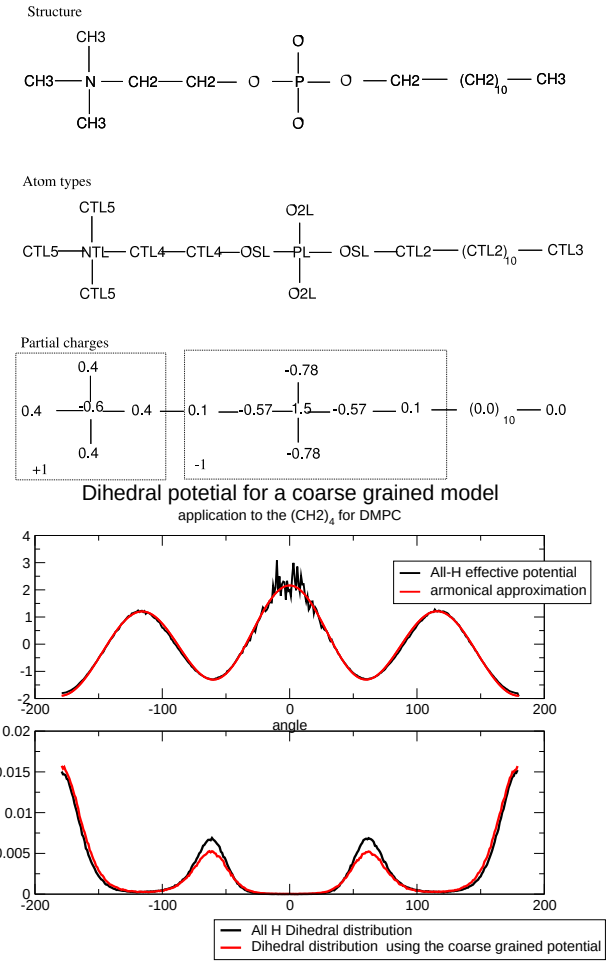


Figure 10: Top: topology of DPC molecule. Bottom: parametrization of the dihedral potential for the united atom model of lipid.

$$S_i^{free} = 4\pi(R_i + r_p)^2$$

The atomic solvent exposed surface S_i is evaluated with the Still's formula.

Polar term The polar surface tension σ_p is extracted by linear fit of computed solvation energies using formula 1 with respect to finite different Poisson electrostatic solvation energies of neutral compounds. The compounds that have been used are Acetyl-X-N-methyl, where X is a non-charged residue, and zwitterionic Phosphatidylcholine (PC). 50 conformations extracted from a simulation of 500 ns at 330 K in vacuum with $\epsilon = 2r$ distance dependent dielectric constant are used for each compound. The value of σ_p obtained from the linear fit (see figure 11) is -96.0.

It is important to note that the polar part of equation 1 is a generalization of the Born formula for the solvation of unbound ions:

$$E_{ion} = \frac{q^2}{R_B} \tag{2}$$

In fact in the ideal case that atom i is completely exposed to the solvent, then $S_i = S_i^{free}$, and $E_p = E_{ion}$ with $q_i = q$ and the Born radius is $R_{Born} = (R_i + r_p)$. In the case that atom i is completely buried, $E_p = 0$. The polar part of equation (1) might be interpreted as the first order approximation of the dielectric descreening spherical integral [50], which after performing the angular integration may be written as [51]:

$$E_p = -\frac{1}{2}(1 - \frac{1}{\epsilon}) \sum_i q_i^2 \int_{\rho_i}^{\infty} \frac{d\rho}{\rho^2} \frac{S_i(\rho)}{4\pi\rho^2} \tag{3}$$

where the ρ_i are tha atoms intrinsic radii. The integral can be written as

$$G(\infty) - G(x) = \int_x^{\infty} \frac{d\rho}{\rho^2} \frac{S_i(\rho)}{4\pi\rho^2} = \int_x^{\infty} F(\rho) d\rho$$

since $G(\infty) = 0$ one obtains that $G'(x) = -F(x)$ and developing $G(x)$ as Maclaurin series around $x = r_i = R_i + r_p$ to the first order:

$$G(x) = G(r_i) + (x - r_i)G'(r_i) = G(r_i) + (r_i - x)F(r_i)$$

which, imposing $x = 2r_i$, reads:

$$G(2r_i) - G(r_i) = -r_i F(r_i)$$

Since $G(2r_i)$ can be neglected with respect to $G(r_i)$ if the charges are small, this brings to the short distance approximation:

$$G(r_i) = r_i F(r_i) \tag{4}$$

that together with the equation (3), the polar term of equation (1) is obtained.

Electrostatic interaction. The electrostatic interaction screening is expressend by the distant dependent approximation $\epsilon(r) = 2r$.

Non-polar term The non-polar term has been obtained using an empirical approach, and two indipendent procedures has been employed for lipids and proteins. A spontaneous aggregation of DPC molecules, and structural investigation of obtained micelle has been adopted as a criterium for choosing the σ_{np} for lipids. 56 initially homogenously dispersed lipids are simulated at 300 K with the Langevin dynamics in a cubic box of 90 $\mathring{A}$. Ten simulation are run in the interval $\sigma_{np} = 0.010 - 0.080 kcal/mol \cdot \mathring{A}^{-2}$. At each value of the non-polar surface ten-

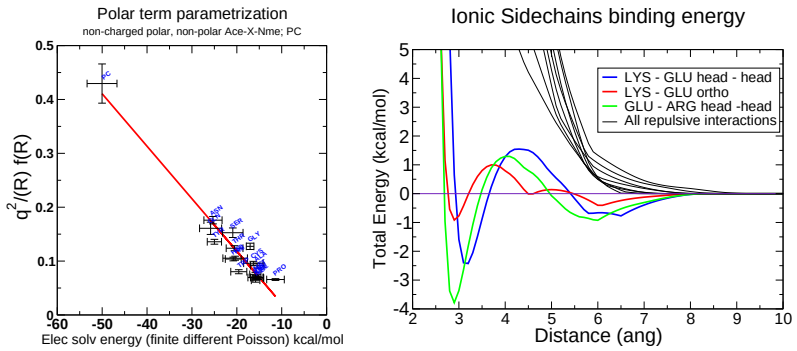


Figure 11: Left: parametrization of the σ_p, i.e. the polar surface tension parameter of equation 1. Right: binding PMF of ionic sidechains analogues.

sion, the radial distribution of heavy atoms with respect to the centre of the self aggregated micelle and the order parameter tensor:

$$S_{ij} = \frac{1}{2}\langle 3\cos\theta_i \cos\theta_j\rangle - \delta_{ij}; i,j \in x,y,z$$

as well as the area per lipid, were evaluated as described in [52]. The σ_{np} value that best approaches the explicit water simulations [52] is 0.043 $kcal/mol \cdot \mathring{A}^{-2}$. Using experimental solvation data of non-polar compounds sigma is 0.025.

Polar solvation for charged groups. A major drawback of using 1 is that the solvation energy of charged residues is underestimated. As an example the direct solvation energy of glutamine is lower than that of glutammate, in contranst with solvation scales (cite some scales). The partial charge sum expressed by formula (1) works correctly for polar groups with neutral charge where the short distance approximation of equation (4) holds, but not for charged chemical moieties. To correct these discrepancies further adjustable parameters are introduced, which are the effective partial charge q_X^* of a residue, where X =LYS, ASP, GLU, ARG and protonated HIS. The different partial charges of atoms belonging to charged chemical groups (such as carboxy for ASP and GLU, guanidinium for ARG and ammonium for LYS) are substituited by the same value q_X^*. To understand the physical meanining of this parameter, let us consider a molecular ion completely

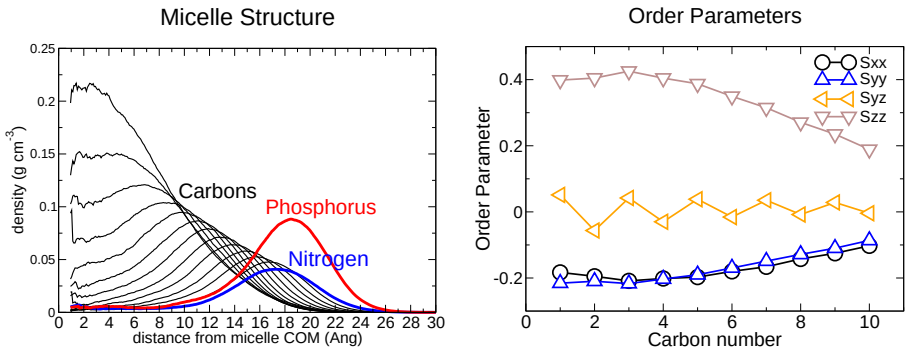


Figure 12: Micelle structure. Left: DPC micelle atomic distribution densities. Right: deuterium order parameter of DPC in micellar aggregates.

exposed to solvent. Using the Born's formula (2) for a molecular system with unitary charge, exploiting the fact that the exposed molecular surface S is equal to the free molecular surface S^{free}:

$$E_p = \frac{1}{R_B} = \frac{1}{R_B}\frac{S}{S^{free}} \tag{5}$$

In the spherical approximation, the molecular surface can be expressed in term of the Born radius $S^{free} = 4/3\pi R_B^3$. The molecular volume is thus defined as $3V_M = R_b \cdot S^{free}$. If $\langle V \rangle = V_M N^{-1}$ is the average atomic partial volume, and using the fact that the sum of surface atomic contribution is the total molecular surface $S = \sum S_i$, then one obtains

$$\frac{1}{R_B} = \frac{1}{3N\langle V \rangle} \cdot \sum_i S_i$$

One can express the average free volume of one atom $\langle V^{free} \rangle$ in term of the average atomic partial volume, by introducing a factor $0 < \lambda < 1$ that expresses the average atomic volume reduction for being buried in a covalent molecular structure:

$$\langle V^{free} \rangle = \frac{4\pi}{3N} \sum_i (R_i + r_p)^3 = \lambda \langle V \rangle \tag{6}$$

Using equation (5) and equation (6) the solvation energy of a molecular ion is thus:

$$E_p = \frac{1}{R_B} = \sum_i \frac{q^{*2}}{r_i + R_P} \cdot \frac{S_i}{S_i^{free}}$$

where the square of the effective partial charge q^{*2} is the factor λ: $q^{*2} = \lambda$. Note that the partial charge is the same for all atoms belonging to the same ion. The polar term for charged sidechains has been parametrized using Potential of Mean force of ionic sidechains analogues [53], see figure 11. We introduce an effective group charge, different from the interaction charge, that appears only in solvation energy.

Transferability of the model The model is not biased toward any particular secondary structure type. In fact, exactly the same force field and implicit solvent

σ_{np}	ARG	LYS	GLU/ASP	Hbond
0.025	0.70	0.90	0.70	-3.36
0.030	0.71	0.91	0.71	-4.09
0.035	0.725	0.925	0.725	-4.85
0.043	0.75	0.95	0.75	-5.99

Table 2: Table of ionic effective charges

model have been used in MD simulations of folding of structured peptides (α-helices and β-sheets) and a number of small proteins.

Kink Angle The kink angle is evaluated whenever helical contacts are formed both in the 1-14 segment and the 15-26 segment. An helix is considered formed when at least two helical turns are present. Given the axis versors v_1 and v_2 the kink angle is $\arccos(v_1 \cdot v_2)$.

Helical contacts An helical contact is defined whenever the backbone amide nitrogen of residue i is close less than 4Å from carbonyl carbon of residue i+4.

References

1. Ladokhin, A. S. & White, S. H. Folding of amphipathic alpha-helices on membranes: energetics of helix formation by melittin. *J Mol Biol* **285**, 1363–1369 (1999).

2. Dawson, C. R., Drake, A. F., Helliwell, J. & Hider, R. C. The interaction of bee melittin with lipid bilayer membranes. *Biochim Biophys Acta* **510**, 75–86 (1978).

3. Dempsey, C. E. The actions of melittin on membranes. *Biochim Biophys Acta* **1031**, 143–161 (1990).

4. Bader, R., Bettio, A., Beck-Sickinger, A. G. & Zerbe, O. Structure and dynamics of micelle-bound neuropeptide Y: comparison with unligated NPY and implications for receptor selection. *J Mol Biol* **305**, 307–329 (2001).

5. Wider, G. Nmr structure of the micelle-bound polypeptide hormone glucagon. *Magnetic Resonance in Chemistry 2003* **41**, S56 (2003).

6. Hsu, S.-T. D. *et al.* NMR study of mersacidin and lipid II interaction in dodecylphosphocholine micelles. Conformational changes are a key to antimicrobial activity. *J Biol Chem* **278**, 13110–13117 (2003).

7. Shenkarev, Z. O. *et al.* Spatial structure of zervamicin IIB bound to DPC micelles: implications for voltage-gating. *Biophys J* **82**, 762–771 (2002).

8. Ladokhin, A. S. & White, S. H. Interfacial folding and membrane insertion of a designed helical peptide. *Biochemistry* **43**, 5782–5791 (2004).

9. Lauterwein, J., Brown, L. R. & Wthrich, K. High-resolution 1h-nmr studies of monomeric melittin in aqueous solution. *Biochim Biophys Acta* **622**, 219–230 (1980).

10. Brown, L. R., Lauterwein, J. & Wüthrich, K. High-resolution 1h-nmr studies of self-aggregation of melittin in aqueous solution. *Biochim Biophys Acta* **622**, 231–244 (1980).

11. Terwilliger, T. C. & Eisenberg, D. The structure of melittin. I. Structure determination and partial refinement. *J Biol Chem* **257**, 6010–6015 (1982).

12. Terwilliger, T. C. & Eisenberg, D. The structure of melittin. II. Interpretation of the structure. *J Biol Chem* **257**, 6016–6022 (1982).

13. Bazzo, R. *et al.* The structure of melittin. a 1h-nmr study in methanol. *Eur J Biochem* **173**, 139–146 (1988).

14. Ikura, T., Go, N. & Inagaki, F. Refined structure of melittin bound to perdeuterated dodecylphosphocholine micelles as studied by 2D-NMR and distance geometry calculation. *Proteins* **9**, 81–89 (1991).

15. Naito, A. *et al.* Conformation and dynamics of melittin bound to magnetically oriented lipid bilayers by solid-state (31)P and (13)C NMR spectroscopy. *Biophys J* **78**, 2405–2417 (2000).

16. Gerig, J. T. Structure and solvation of melittin in 1,1,1,3,3,3-hexafluoro-2-propanol/water. *Biophys J* **86**, 3166–3175 (2004).

17. Gerig, J. T. Structure and solvation of melittin in hexafluoroacetone/water. *Biopolymers* **74**, 240–247 (2004).

18. Dempsey, C. E. *et al.* Contribution of proline-14 to the structure and actions of melittin. *FEBS Lett* **281**, 240–244 (1991).

19. Hewish, D. R. *et al.* Structure and activity of D-Pro14 melittin. *J Protein Chem* **21**, 243–253 (2002).

20. Berneche, S. & Roux, B. Energetics of ion conduction through the K+ channel. *Nature* **414**, 73–77 (2001).

21. de Groot, B. L. & Grubmüller, H. Water permeation across biological membranes: mechanism and dynamics of aquaporin-1 and GlpF. *Science* **294**, 2353–2357 (2001).

22. Tajkhorshid, E. *et al.* Control of the selectivity of the aquaporin water channel family by global orientational tuning. *Science* **296**, 525–530 (2002).

23. Böckmann, R. A. & Caflisch, A. Spontaneous formation of detergent micelles around the outer membrane protein OmpX. *Biophys J* **88**, 3191–3204 (2005).

24. Gorfe, A. A., Pellarin, R. & Caflisch, A. Membrane localization and flexibility of a lipidated ras peptide studied by molecular dynamics simulations. *J Am Chem Soc* **126**, 15277–15286 (2004).

25. Lagüe, P., Roux, B. & Pastor, R. W. Molecular dynamics simulations of the influenza hemagglutinin fusion peptide in micelles and bilayers: conformational analysis of peptide and lipids. *J Mol Biol* **354**, 1129–1141 (2005).

26. Glättli, A., Chandrasekhar, I. & van Gunsteren, W. F. A molecular dynamics study of the bee venom melittin in aqueous solution, in methanol, and inserted in a phospholipid bilayer. *Eur Biophys J* **35**, 255–267 (2006).

27. Berneche, S., Nina, M. & Roux, B. Molecular dynamics simulation of melittin in a dimyristoylphosphatidylcholine bilayer membrane. *Biophys J* **75**, 1603–1618 (1998).

28. Bachar, M. & Becker, O. M. Protein-induced membrane disorder: a molecular dynamics study of melittin in a dipalmitoylphosphatidylcholine bilayer. *Biophys J* **78**, 1359–1375 (2000).

29. Lin, J. H. & Baumgaertner, A. Stability of a melittin pore in a lipid bilayer: a molecular dynamics study. *Biophys J* **78**, 1714–1724 (2000).

30. Lazaridis, T. Implicit solvent simulations of peptide interactions with anionic lipid membranes. *Proteins* **58**, 518–527 (2005).

31. Spassov, V., Yan, L. & Szalma, S. Introducing an implicit membrane in generalized born/solvent accessibility continuum solvent models. *J. Phys. Chem. B* **106**, 8726 – 8738 (2002).

32. Im, W. & Brooks, C. L. Interfacial folding and membrane insertion of designed peptides studied by molecular dynamics simulations. *Proc Natl Acad Sci U S A* **102**, 6771–6776 (2005).

33. Maddox, M. W. & Longo, M. L. A monte carlo study of peptide insertion into lipid bilayers: equilibrium conformations and insertion mechanisms. *Biophys J* **82**, 244–263 (2002).

34. Yau, W. M., Wimley, W. C., Gawrisch, K. & White, S. H. The preference of tryptophan for membrane interfaces. *Biochemistry* **37**, 14713–14718 (1998).

35. Dempsey, C. E. & Butler, G. S. Helical structure and orientation of melittin in dispersed phospholipid membranes from amide exchange analysis in situ. *Biochemistry* **31**, 11973–11977 (1992).

36. Chen, X., Wang, J., Boughton, A. P., Kristalyn, C. B. & Chen, Z. Multiple orientation of melittin inside a single lipid bilayer determined by combined vibrational spectroscopic studies. *J Am Chem Soc* **129**, 1420–1427 (2007).

37. Brown, L. R., Braun, W., Kumar, A. & Wüthrich, K. High resolution nuclear magnetic resonance studies of the conformation and orientation of melittin bound to a lipid-water interface. *Biophys J* **37**, 319–328 (1982).

38. Lauterwein, J., Bösch, C., Brown, L. R. & Wthrich, K. Physicochemical studies of the protein-lipid interactions in melittin-containing micelles. *Biochim Biophys Acta* **556**, 244–264 (1979).

39. Inagaki, F. *et al.* Structure of melittin bound to perdeuterated dodecylphosphocholine micelles as studied dy two-dimensional nmr and distance geometry calculations. *Biochemistry* **28**, 5985–5991 (1989).

40. Neidigh, J. W. & Andersen, N. H. Peptide conformational changes induced by tryptophan-phosphocholine interactions in a micelle. *Biopolymers* **65**, 354–361 (2002).

41. de Planque, M. R. *et al.* Different membrane anchoring positions of tryptophan and lysine in synthetic transmembrane alpha-helical peptides. *J Biol Chem* **274**, 20839–20846 (1999).

42. Schiffer, M., Chang, C. H. & Stevens, F. J. The functions of tryptophan residues in membrane proteins. *Protein Eng* **5**, 213–214 (1992).

43. Kaznessis, Y. N., Kim, S. & Larson, R. G. Specific mode of interaction between components of model pulmonary surfactants using computer simulations. *J Mol Biol* **322**, 569–582 (2002).

44. Freites, J. A., Choi, Y. & Tobias, D. J. Molecular dynamics simulations of a pulmonary surfactant protein B peptide in a lipid monolayer. *Biophys J* **84**, 2169–2180 (2003).

45. Altenbach, C., Froncisz, W., Hyde, J. S. & Hubbell, W. L. Conformation of spin-labeled melittin at membrane surfaces investigated by pulse saturation recovery and continuous wave power saturation electron paramagnetic resonance. *Biophys J* **56**, 1183–1191 (1989).

46. Brooks, B. R. *et al.* CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comp. Chem.* **4**, 187–217 (1983).

47. Rao, F. & Caflisch, A. Replica exchange molecular dynamics simulations of reversible folding. *Journal of Chemical Physics* **119**, 4035–4042 (2003).

48. Feller, S. & MacKerell, A. An improved empirical potential energy function for molecular simulations of phospholipids. *Journal of Physical Chemistry B* **104**, 7510–7515 (2000).

49. Egberts, E., Marrink, S. J. & Berendsen, H. J. Molecular dynamics simulation of a phospholipid membrane. *Eur Biophys J* **22**, 423–436 (1994).

50. Still, W., Tempczyk, A., Hawley, R. C. & Hendrickson, T. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* **112**, 6127–6129 (1990).

51. Hawkins, G. D., Cramer, C. J. & Truhlar, D. G. Pairwise solute descreening of solute charges from a dielectric medium. *Chemical Physics Letters* **246**, 122–129 (1995).

52. Marrink, S., Tieleman, D. & Mark, A. Molecular dynamics simulation of the kinetics of spontaneous micelle formation. *J. Phys. Chem. B* **104**, 12165–12173 (2000).

53. Masunov, A. & Lazaridis, T. Potentials of mean force between ionizable amino acid side chains in water. *J Am Chem Soc* **125**, 1722–1730 (2003).

9 Explicit solvent simulation of a lipidated peptide

Among all the membrane associated peptides, lipid-modified proteins binds to the membrane by inserting their lipid chain into the hydrophobic core. These proteins plays important roles in the events of signal transduction of the cell. A particular noticeable example is the GTPase ras, which is involved in cell proliferation [92]. A detailed knowledge of the ras signal transmission is fundamental, since about one-third of human cancers present a mutated form of ras proteins. The C-terminal segment GTPase ras contains two lipidated cysteines: since downstream effectors occur at the membrane surface, post-translational lipidation of these two residues regulates the protein function by partitioning the C-terminal portion at the membrane surface. Spectroscopic techniques were used to investigate the membrane localization of the heptapeptide 180-186 [93]. The authors gave evidence that lipid chains were deeply inserted into the membrane, that hydrophobic amino acid side chains were partitioned at the membrane interior, and the peptide backbone is disordered and preferentially resides at the membrane-water interface. In this work (see section 9.1) several explicit water MD simulations have been performed, to investigate the stability, the conformations and the interaction of the membrane-peptide system (see figure 7). The position and orientation of the ras peptide and its components obtained by MD simulations are consistent with spectroscopic data [93], and a mechanism of insertion is proposed.

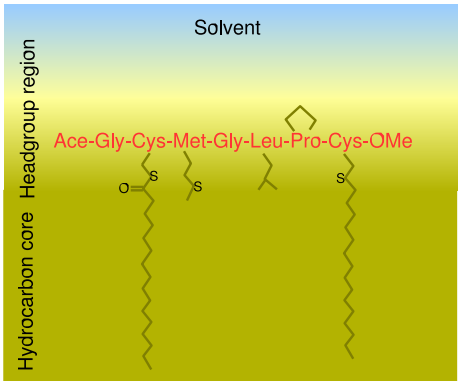


Figure 7: Schematic representation of ras peptide partitioning at the different membrane environments, as resulted from simulation analysis presented in section 9.1.

9.1 Membrane localization and flexibility of a lipidated ras peptide studied by molecular dynamics simulations.[J. Am. Chem. Soc. 2004, *126*, 15277]

J|A|C|S
ARTICLES
Published on Web 11/02/2004

Membrane Localization and Flexibility of a Lipidated Ras Peptide Studied by Molecular Dynamics Simulations

Alemayehu A. Gorfe, Riccardo Pellarin, and Amedeo Caflisch*

Contribution from the Department of Biochemistry, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland

Received June 9, 2004; E-mail: caflisch@bioc.unizh.ch

Abstract: Lipid-modified membrane-binding proteins are essential in signal transduction events of the cell, a typical example being the GTPase ras. Recently, membrane binding of a doubly lipid-modified heptapeptide from the C-terminus of the human N-ras protein was studied by spectroscopic techniques.[14] It was found that membrane binding is mainly due to lipid chain insertion, but it is also favored by interactions between apolar side chains and the hydrophobic region of the membrane. Here, 10 explicit solvent molecular dynamics simulations for a total time of about 150 ns are used to investigate the atomic details of the peptide−membrane association. The 16:0 peptide lipid chains are more mobile than the 14:0 phospholipid chains, which is in agreement with ^{2}H NMR experiments. Peptide−lipid and peptide−solvent interactions, backbone and side-chain distributions, as well as the effects of lipidated peptide insertion onto the structure, and dynamics of a 1,2-dimyristoylglycero-3-phosphocholine bilayer are described. The simulation results validate the structural model proposed by the analysis of spectroscopic data and highlight the main aspects of the insertion mechanism. The peptide in the membrane is rather rigid over the simulation time scale of about 10 ns, but different partially extended conformations devoid of backbone hydrogen bonds are observed in different trajectories.

Introduction

The regulation of cellular functions is coordinated by signal molecules. External signals transmitted across membranes and received by cellular receptors are relayed to their target by intracellular signal cascades. Posttranslational lipid-modified proteins[1,2] are commonly involved in regulation of the signal transmission processes.[3−5] Typical among fatty-acid-modified proteins is the GTPase ras. The ras signal transduction cascade is central in cell proliferation and differentiation events, in which ras proteins, by activating downstream effectors,[6,7] mediate the signal flow from receptor tyrosine kinase to the cell nucleus. Malfunction in the regulatory action (the switching function) of ras proteins leads to uncontrolled cell growth, or cancer, manifested by the fact that about one-third of all human cancers carry a mutated form of ras proteins. Detailed characterization of the molecular interactions playing a role in the ras signal transmission pathways is therefore of great importance.

The specific attachment of lipids to proteins that contain the C-terminal motif, CAAX, by cleavage and addition to the SH group of Cys, is a common structural feature in membrane proteins involved in signal transduction events. Lipidated proteins (and their analogue peptides) achieve strong binding by inserting their lipid chains into the hydrophobic core of the membrane. Thus, the major membrane-binding energy contribution comes from the hydrophobic interaction between the hydrocarbon tails of the protein and the membrane. For a maximum binding potential, ras proteins can acquire two different types of lipid modification, single or double lipid modification. Typically, ras proteins with single lipid modification also contain cluster(s) of basic amino acids and bind to negatively charged plasma membranes. The interaction between the phospholipid headgroups (of the negatively charged membranes) and the positive peptide charges provides additional attractive electrostatic energy for a stable membrane association (e.g., the K-ras protein[8]). H- and N-ras proteins require double modifications.[9] Farnesylated, but nonpalmitoylated, H- and N-ras proteins mislocate to the cytosol and break the signal cascade.[10−12] The human N-ras protein, the focus of this paper, undergoes farnesylation at the C-terminal recognition region (Cys186) followed by palmitoylation at Cys181. Single lipid modification of N-ras provides insufficient hydrophobic binding energy for it to permanently anchor to plasma membranes. As a result, a fast equilibrium between adsorbed and desorbed states is observed.

(1) Hancock, J. F.; Cadwallader, K.; Paterson, H.; Marshall, C. J. *EMBO J.* **1991**, *10*, 4033−4039.
(2) Hancock, J. F.; Paterson, H.; Marshall, C. J. *Cell* **1990**, *63*, 133−139.
(3) Reuther, G. W.; Der, C. J. *Curr. Opin. Cell Biol.* **2000**, *12*, 157−165.
(4) Miggin, S. M.; Lawler, O. A.; Kinsella, B. T. *J. Biol. Chem.* **2003**, *278*, 6947−6958.
(5) Clarke, S. *Annu. Rev. Biochem.* **1992**, *61*, 355−386.
(6) Scheffzek, K.; Ahmadian, M. R.; Kabsch, W.; Wiesmuller, L.; Lautwein, A.; Schmitz, F.; Wittinghofer, A. *Science* **1997**, *277*, 333−338.
(7) Boguski, M. S.; McCormick, F. *Nature* **1993**, *366*, 643−654.
(8) Ghomashchi, F.; Zhang, X.; Liu, L.; Gelb, M. H. *Biochemistry* **1995**, *34*, 11910−11918.
(9) Silvius, J. R. Lipidated Peptides as Tools for Understanding the Membrane Interactions of Lipid-Modified Proteins. In *Peptide Lipid*; Simon, S. A., McIntosh, T. J., Eds.; Elsevier: New York, 2002; pp 371−395.
(10) Peters, C.; Wagner, M.; Volkert, M.; Waldmann, H. *Naturwissenschaften* **2002**, *89*, 381−390.
(11) Dudler, T.; Gelb, M. H. *J. Biol. Chem.* **1996**, *271*, 11541−11547.

10.1021/ja046607n CCC: $27.50

Unlike integral proteins that are permanently anchored to membranes, the association of ras proteins to plasma membranes is an equilibrium process.[13] Because interactions with other downstream effectors occur at the membrane surface, ras proteins are functional only in the membrane-associated state and are inactive desorbed into the cytosol. Experimental studies on membrane interactions of lipid-modified proteins, mainly using lipidated peptides and artificial membranes,[9,13] have shed light on how their distribution is regulated between the active (membrane-bound) and inactive (unbound) states, as well as on the energetics and thermodynamics of the interaction.

Using a combination of Fourier transform infrared, solid-state NMR, and neutron diffraction spectroscopy, Huster et al. recently studied membrane insertion and localization of a heptapeptide representing the carboxy terminus (residues 180−186) of the human N-ras protein.[14] The plasma membrane was modeled by 1,2-dimyristoylglycero-3-phosphocholine (DMPC). This study highlighted the key structural features that accompany membrane insertion of ras proteins. The peptide inserts its two lipid chains deep into the membrane interior, such that stabilizing hydrophobic contacts between the DMPC and peptide lipid chains are possible. Binding is further assisted by the insertion of two hydrophobic amino acid side chains (Leu and Met) into the hydrophobic section of the membrane. The backbone adopts a disordered conformation and is preferentially localized in the lipid−water interface. These observations led to a plausible structural model for membrane binding of N-ras proteins.

It is worth noting that while the structure of the soluble part of ras proteins (residues 1−166) has been solved by both X-ray diffraction[15−18] and solution NMR spectroscopy,[19] there are no structural data for the C-terminal membrane binding region (residues 180−186). The biophysically derived structural model of Huster et al. for a doubly lipid-modified synthetic heptapeptide bound to a DMPC bilayer mimics the binding mode of the human N-ras protein. It is therefore of general significance. However, not all of the details of association can be observed spectroscopically; atomic level analysis of peptide−lipid and peptide−solvent interactions, backbone and side-chain distributions, as well as the effect of lipidated-peptide insertion onto the structure, and dynamics of a DMPC bilayer are required for a molecular-level interpretation and full description of the association process. A more-direct approach to investigate these and related issues is provided by computational methods. Due to the level of details that they can provide, molecular dynamics (MD) simulations of the lipidated peptide and DMPC bilayer are uniquely suited to address these questions at the atomic level.

Several explicit water MD studies of membrane−protein systems have been conducted (see, for example, review articles[20−22] and recent reports[23−25]). The majority of these simulations targeted the stability, dynamics, and functional aspects of integral proteins in a variety of membranes,[26−32] pure membranes and their environments,[33−35] and peptide−bilayer systems.[36−40] So far, however, only few attempts have been made to simulate membrane localization, and the accompanying mechanisms of insertion of peptides/proteins whose initial positions, with respect to the phospholipids, are not clearly known a priori.[41−43] Kaznessis et al. investigated the interaction between the N-terminal region of the human surfactant protein-B in dipalmitoylphosphatidylcholine (DPPC) and dipalmitoylphosphatidylglycerol (DPPG) monolayers.[41] Their simulations revealed that the peptide fragment of protein-B adopts different modes and energetics of interactions with the different phospholipid monolayers, with preferential affinity for anionic phospholipids. Knecht and Grubmüller used MD and annealing simulations to study mechanical coupling by the membrane fusion SNARE protein syntaxin 1A.[42] Their simulation results indicate that partially unstructured linkers provide significant mechanical coupling. Sankararamakrishnan and Weinstein simulated the helical region of dynorphin in a DMPC bilayer.[43] They showed that in the complex, the tilt angle of the dynorphin helix from the bilayer normal is stabilized at ∼50° for different initial orientations of the dynorphin.[43] In a recent study on the insertion of antimicrobial peptides into a zwitterionic lipid bilayer in which the peptides were directed toward the interface either on their hydrophobic or positively charged face, it was found that the former bind to the interface and subsequently penetrate the bilayer while the latter displayed only partial surface binding.[44]

In the present study, MD simulations are used to investigate the position, orientation, and flexibility of a lipidated ras peptide

(12) Nägele, E.; Schelhaas, M.; Kunder, N.; Waldmann, H. *J. Am. Chem. Soc.* **1998**, *120*, 6889−6902.
(13) Hinterding, K.; Alonso-Diaz, D.; Waldmann, H. *Angew. Chem., Int. Ed.* **1998**, *37*, 688−749.
(14) Huster, D.; Vogel, A.; Katzeka, C.; Sheidt, H. A.; Binder, H.; Dante, S.; Gutberlet, T.; Schoernig, O.; Waldmann, H.; Arnold, K. *J. Am. Chem. Soc.* **2003**, *125*, 4070−4079.
(15) Pai, E. F.; Kabsch, W.; Krengel, U.; Holmes, K. C.; John, J.; Wittinghofer, A. *Nature* **1989**, *341*, 209−214.
(16) Milburn, M. V.; Tong, L.; deVos, A. M.; Brunger, A.; Yamaizumi, Z.; Nishimura, S.; Kim, S. H. *Science* **1990**, *247*, 939−945.
(17) Brunger, A. T.; Milburn, M. V.; Tong, L.; deVos, A. M.; Jancarik, J.; Yamaizumi, Z.; Nishimura, S.; Ohtsuka, E.; Kim, S. H. *Proc. Natl. Acad. Sci. U.S.A.* **1990**, *87*, 4849−4853.
(18) Krengel, U.; Schlichting, L.; Scherer, A.; Schumann, R.; Frech, M.; John, J.; Kabsch, W.; Pai, E. F.; Wittinghofer, A. *Cell* **1990**, *62*, 539−548.
(19) Kraulis, P. J.; Domaille, P. J.; Campbell-Burk, S. L.; Akenn, T. V.; Laue, E. D. *Biochemistry* **1994**, *33*, 3515−3531.
(20) Pastor, R. W.; Venable, R. M.; Feller, S. E. *Acc. Chem. Res.* **2002**, *35*, 438−446.
(21) Faraldo-Gomez, J. D.; Smith, G. R.; Sansom, M. S. *Eur. Biophys. J.* **2002**, *31*, 217−227.
(22) Liang, J. *Curr. Opin. Chem. Biol.* **2002**, *6*, 878−884.
(23) Aksimentiev, A.; Balabin, I. A.; Fillingame, R. H.; Schulten, K. *Biophys. J.* **2004**, *86*, 1332−1344.
(24) Allen, T. W.; Andersen, O. S.; Roux, B. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 117−122.
(25) Gao, M.; Craig, D.; Lequin, O.; Campbell, I. D.; Vogel, V.; Schulten, K. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 14784−14789.
(26) Domene, C.; Bond, P. J.; Sansom, M. S. *Adv. Protein Chem.* **2003**, *66*, 159−193.
(27) Saiz, L.; Bandyopadhyay, S.; Klein, M. L. *Biosci. Rep.* **2002**, *22*, 151−173.
(28) Sansom, M. S.; Shrivastava, I. H.; Bright, J. N.; Tate, J.; Capener, C. E.; Biggin, P. C. *Biochim. Biophys. Acta* **2002**, *1565*, 294−307.
(29) Chung, S. H.; Kuyucak, S. *Biochim. Biophys. Acta* **2002**, *1565*, 267−286.
(30) Hansson, T.; Oostenbrink, C.; van Gunsteren, W. *Curr. Opin. Struct. Biol.* **2002**, *12*, 190−196.
(31) Sansom, M. S.; Bond, P.; Beckstein, O.; Biggin, P. C.; Faraldo-Gomez, J.; Law, R. J.; Patargias, G.; Tieleman, D. P. *Novartis Found. Symp.* **2002**, *245*, 66−78; discussion 79−83, 165−168.
(32) Roux, B. *Acc. Chem. Res.* **2002**, *35*, 366−375.
(33) Róg, T.; Pasenkiewicz-Gierula, M. *Biophys. J.* **2002**, *81*, 2190−2202.
(34) Böckmann, R. A.; Grubmüller, H. *Angew. Chem., Int. Ed.* **2004**, *43*, 1021−1024.
(35) Böckmann, R. A.; Hac, A.; Heimburg, T.; Grubmüller, H. *Biophys. J.* **2003**, *85*, 1647−1655.
(36) La Rocca, P.; Biggin, P. C.; Tieleman, D. P.; Sansom, M. S. *Biochim. Biophys. Acta* **1999**, *1462*, 185−200.
(37) Aliste, M. P.; MacCallum, J. L.; Tieleman, D. P. *Biochemistry* **2003**, *42*, 8976−8987.
(38) Law, R. J.; Tieleman, D. P.; Sansom, M. S. *Biophys. J.* **2003**, *84*, 14−27.
(39) Tarek, M.; Maigret, B.; Chipot, C. *Biophys. J.* **2003**, *85*, 2287−2298.
(40) Tieleman, D. P.; Sansom, M. S.; Berendsen, H. J. *Biophys. J.* **1999**, *76*, 40−49.
(41) Kaznessis, Y. N.; Kim, S.; Larson, R. G. *J. Mol. Biol.* **2002**, *322*, 569−582.
(42) Knecht, V.; Grubmüller, H. *Biophys. J.* **2003**, *84*, 1527−1547.
(43) Sankararamakrishnan, R.; Weinstein, H. *Biophys. J.* **2000**, *79*, 2331−2344.
(44) Shepherd, C. M.; Vogel, H. J.; Tieleman, D. P. *Biochem. J.* **2003**, *370*, 233−243.

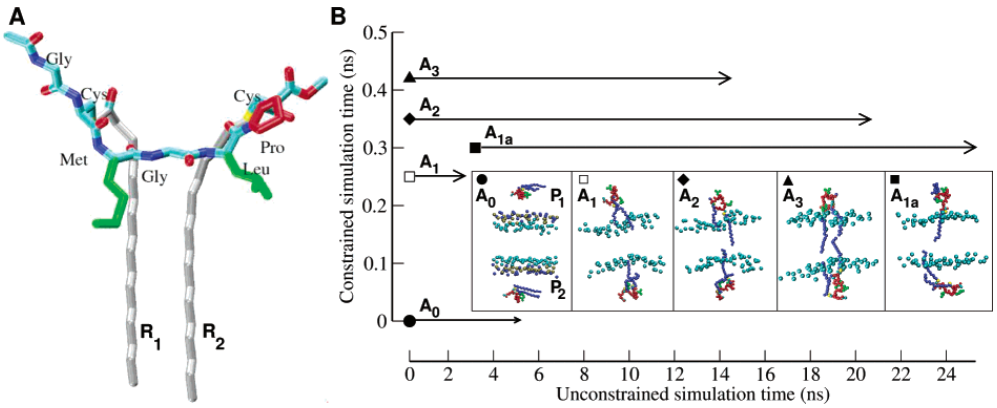


Figure 1. (A) Sequence and starting structure of the acetylated form of the N-ras peptide used in the simulations. The structure is modeled as described in the text. (B) Schematic picture of the approach used to determine initial peptide positions with respect to the bilayer. After construction of the bilayer−water system (see text), one peptide was placed in the middle of the water slabs at each monolayer. The system was relaxed for 0.25 ns. Then, one simulation was continued without any bias (A_0), while in another simulation, the separation along the membrane normal between the methyl carbon atoms of the ras lipids [C_{16}, $R_1(P_1)$ and C_{16}, $R_1(P_2)$; and C_{16}, $R_2(P_1)$ and C_{16}, $R_2(P_2)$] was decremented by 0.5 Å every 5 ps using a harmonic potential with a force constant of 100 kcal mol^{-1} Å^{-2}. The constraint was removed after 0.25 ns (starting structure for A_1), 0.35 ns (A_2), and 0.42 ns (A_3). Since P_2 and P_1 in trajectories A_1 and C_1 have completely dissociated to solvent (see Results and Discussion), the last snapshots of these were used to start trajectories A_{1a} and C_{1a}, respectively. This was done by "pulling-back-in" one chain of the dissociated peptide into the bilayer. Vertical and horizontal time axes indicate the length of the simulations with and without a harmonic constraint, respectively. The five insets show the peptide positions and orientations used as starting structures for production runs. Ras lipid chains are in blue; backbone and residue prolines are in red, while Leu and Met are in green. For clarity, only selected atoms of the DMPC are shown: phosphorus and choline nitrogen atoms in dark yellow and blue, respectively, and the first methylene carbon of the lipid tails in cyan. A similar approach was followed to obtain initial positions for simulations with a charged N-terminus.

Table 1. Performed Simulations[a]

	Acetylated N-Terminus					Charged N-Terminus				
name	A_0	A_1	A_2	A_3	A_{1a}	C_0	C_1	C_2	C_3	C_{1a}
constrained dynamics (ns)	0.0	0.25	0.35	0.42	0.10[b]	0.0	0.35	0.40	0.45	0.10[b]
unconstrained dynamics (ns)	5.0	2.3	20.1	14.1	22.5	5.0	2.6	22.3	14.2	20.0

[a] See Figure 1B for details. [b] Simulations were continued from simulations A_1 and C_1, and the constraints were applied on a single lipid chain (see text).

in a DMPC bilayer. The simulation results are first validated by comparison with NMR, Fourier transform infrared, and neutron diffraction spectroscopy experiments.[14] The good agreement between simulation and experimental results allows the use of the former to extract a clear and detailed picture of membrane insertion of the lipidated ras peptide.

Methods

It was shown experimentally that equilibrium membrane adsorption of doubly lipid-modified ras peptides have average half-life times in the order of hours to days.[45,46] Although it would be interesting to simulate the system long enough for the peptide to insert spontaneously, this is impossible to achieve within the current accessible time scales of MD simulations. A procedure to speed up the insertion is therefore necessary. The initial position and orientation of the peptide with respect to the bilayer were chosen such that insertion could be observed within a reasonable simulation time. Furthermore, to maximize sampling, two peptides (i.e., one peptide per leaflet) were simulated in each MD run. This is partly justified by the fact that the lipid bilayers are symmetrical in a simulation box and also because, in the in vitro experiments, the two leaflets were both populated.

Peptide Structure. The peptide sequence used in the simulations consisted of residues (X)Gly-Cys(R_1)-Met-Gly-Leu-Pro-Cys(R_2)-OMe, where R_1 and R_2 represent the lipid modifications (i.e., palmitic and hexadecyl thioether tail, respectively).[14] The latter was preferred to the natural occurring farnesyl to directly compare with the experimental data.[14] The X represents a hydrogen atom in the case of a charged N-terminus and an acetyl (CH_3OC-) in the neutral form of the peptide. Since the peptide sequence corresponds to the C-terminal end of the N-ras protein, the neutral form is closer to the natural protein, while the charged form corresponds to the one used in the experiments.[14] Parameters for the lipid modifications were derived from the CHARMM27 force field.[47,48] A model for the structure of the peptide was then built manually and minimized by 1000 steps SD and 1000 steps conjugate gradient to remove bad atomic contacts, leading to the structure shown in Figure 1A.

Initial Peptide Positions and Orientations. Figure 1B and Table 1 schematically show the strategy used to obtain starting structures. First, the peptides were placed in the middle of each water slab and relaxed for 0.25 ns. Then, one trajectory was continued without any bias (A_0), and another trajectory was continued with distance constraints. In the latter, the distances along the *z*-axis between corresponding C_{16} atoms of peptide 1 (P_1) and peptide 2 (P_2), i.e., $R_1(P_1)-R_1(P_2)$ and $R_2(P_1)-R_2(P_2)$, were slowly decreased by applying a harmonic force constant (see legend of Figure 1B for details). By removing the

(45) Shahinian, S.; Silvius, J. R. *Biochemistry* **1995**, *34*, 3813−3822.
(46) Schroeder, H.; Leventis, R.; Rex, S.; Schelhaas, M.; Nagele, E.; Waldmann, H.; Silvius, J. R. *Biochemistry* **1997**, *36*, 13102−13109.
(47) Feller, S. E.; Yin, D.; Pastor, R. W.; MacKerell, A. D., Jr. *Biophys. J.* **1997**, *73*, 2269−2279.
(48) MacKerell, A. D., Jr.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Kuchnir, L.; Guo, L.; Guo, H.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Joseph-McCarthy, D.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E., III.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stone, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Karplus, M. *J. Phys. Chem. B* **1998**, *102*, 3586−3616.

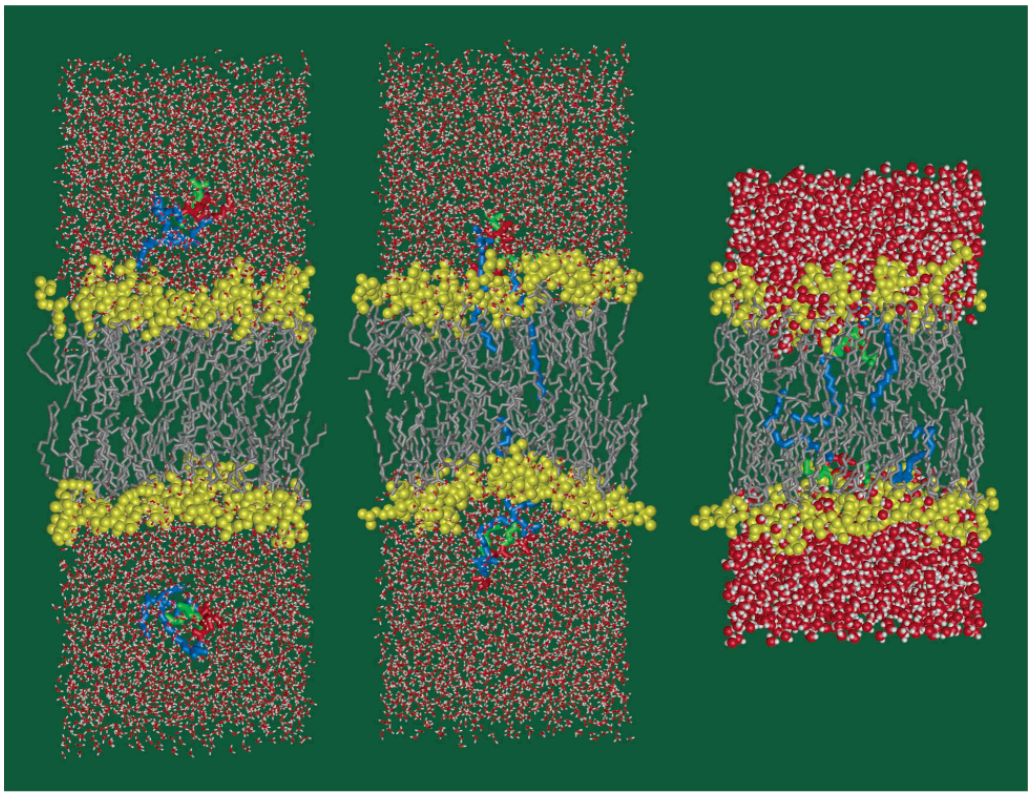


Figure 2. Snapshots after 0.25 ns of relaxation (A_0, left), after partial insertion promoted by a 0.35 ns constrained dynamics (A_2, middle), and the final snapshot (A_2, right). Backbone and Pro are in red, ras lipid chains in blue, Leu and Met side chains in green, DMPC hydrocarbon tails in gray, headgroup atoms in yellow, and water molecules in sticks (or ball-and-sticks). Part of the water molecules have been removed after partial insertion of the peptide. Hydrogen atoms of the peptide and the DMPCs are omitted for clarity.

constraint at 0.25 ns, 0.35 ns, and 0.42 ns, we obtained three starting configurations for simulations A_1, A_2, and A_3. A similar procedure was used to generate starting structures for another four trajectories calculated with a charged N-terminus (C_0-C_3). Furthermore, two trajectories, A_{1a} and C_{1a}, were started from the final conformation of simulations A_1 and C_1, respectively (see legend of Figure 1B). In total, 10 simulations were run to investigate the insertion process (Table 1).

Simulation Protocol. The CHARMM program[49] and the CHARMM27[47,48] parameters were used in all of the simulations and for part of the analysis. The construction and setup of the simulation system were based on the protocol of Woolf and Roux,[50,51] adjusted to the specificities of the present system. The liquid-crystalline phase of the DMPC (at a temperature of 310 K, the temperature at which most of the experiments were done[14]) was simulated at a constant number of particles (N), normal pressure (P_N), cross sectional area (A), and temperature (T), the NP_NAT ensemble.

To avoid strain of the bilayer due to the insertion of the two "foreign" lipid chains of the ras peptide, the cross sectional area of each monolayer was made slightly larger than that in a pure DMPC bilayer. Thus, the total lateral area per leaflet was calculated for 27 lipids with an area per lipid of 59.8 Å^2 (ref 52) to be used for 26 lipids per leaflet in all of the simulations involving peptide insertion. The peptide:DMPC ratio was therefore 1:26 in the simulations, in contrast with the experimental ratio of 1:10. Note that a more-dilute solution is preferable to make sure that multimer formation is avoided, as was also mentioned by Huster et al.. In the control simulation of a DMPC bilayer without the peptides, 27 lipids per leaflet were used within the same cross-sectional area.

The DMPC bilayer was constructed by randomly choosing structures from the pre-equilibrated phospholipid structural libraries.[53,54] To model the bulk solvent, a water slab was constructed from a pre-equilibrated TIP3 model and overlayed on the glycerol region of each leaflet. One peptide was then inserted into each of the water slabs, and water molecules closer than 2.6 Å to any peptide atom were deleted. The size of the water box was chosen such that any atom of a peptide is at least 10 Å away from the edge of the box, including the side parallel to the lipid lateral surface. The dimension of the system was 43.6 × 37.2 × 120.0 Å^3, resulting in a total of ∼4350 water molecules, 2 peptides, and 52 DMPC lipids. In the case of the simulations with a charged N-terminus, the system was neutralized by adding a chloride ion in each water box. The simulation systems for the neutral and charged peptides contained a total of 19 560 and 19 575 atoms, respectively. Figure 2 (left and middle) shows representative snapshots of the simulation setup.

(49) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. T.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187−217.
(50) Woolf, T. B.; Roux, B. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 11631−11635.
(51) Woolf, T. B.; Roux, B. *Proteins* **1996**, *24*, 92−114.
(52) Petrache, H. I.; Dodd, S. W.; Brown, M. F. *Biophys. J.* **2000**, *79*, 3172−3192.
(53) Venable, R. M.; Zhang, Y.; Hardy, B. J.; Pastor, R. W. *Science* **1993**, *262*, 223−226.
(54) de Loof, H. D.; Harvey, S. C.; Segrest, J. P.; Pastor, R. W. *Biochemistry* **1991**, *30*, 2099−2113.

Subsequent minimizations and a 200 ps equilibration with progressively decreasing harmonic constraints on the peptide backbone, side chains, lipid headgroups, and water oxygen atoms were similar to those of previous reports.[55] After equilibration, production simulations were run in the NP_NAT ensemble. Periodic boundary conditions in all three spatial directions were used, with constant normal pressure ($P_N = 1$ atm) and constant A and T. Constant temperature was maintained using the Hoover temperature control[56] with a thermal piston mass of 3000 kcal mol^{-1} ps^2. Truncation of electrostatic interactions has been shown to have major effects on the bilayer properties.[57] In this work, long-range electrostatic interactions were treated by the particle mesh Ewald (PME) method[58] with a 12 Å cutoff for direct and reciprocal space summations. A shift function at 10 Å for the Lennard-Jones interactions and a heuristic update of the nonbonded list, with a cutoff at 12 Å, were used. The integration time step was 2 fs, and all bonds involving hydrogens were fixed using the SHAKE algorithm. Structures were saved every 1 ps for analysis.

To speed up sampling, part of the water layer was removed after a portion of the peptides had inserted into the bilayer, while keeping sufficient waters to solvate both the peptides and the bilayer. This resulted in a reduction of 40% in the total number of atoms (Figure 2, right). In the early stages of the simulations, drift of the peptide along the lateral spatial directions (i.e., perpendicular to the membrane normal) was prevented by applying a cylindrical potential.[50] This constraint was removed after partial insertion to allow a spontaneous lateral reorganization of the peptide in the bilayer.

Analysis. The bilayer thickness (D_{PP}) is defined as the distance between the geometric center of the phosphorus atoms at each monolayer. The average chain length (L_C) is defined as the average distance along the bilayer normal between the first methylene carbon and the terminal methyl carbon atoms of the lipid chains,[59] calculated from trajectories as

$$\langle L_C \rangle \equiv \langle z_2 \rangle - \langle z_x \rangle \tag{1}$$

where x is 14 for the DMPC chains and 16 for the ras chains, and the $\langle\rangle$ denote time averages. Note that L_C, defined here, is equivalent to L^*_C in the literature,[59] which does not include the distance from the first methylene to the carbonyl carbon (~0.55 Å) and the extra length of the terminal methyl group (~0.98 Å).[59]

The deuterium order parameter, S_{CD}, was calculated as

$$S_{CD} = \frac{1}{2}\langle 3\cos^2\theta_n - 1\rangle \tag{2}$$

where θ_n is the instantaneous angle between a vector along the methylene/methyl hydrogens of the acyl carbon atoms and the bilayer normal.

Graphical analysis of the simulations was made using the VMD program.[60]

Results and Discussion

Control Simulations. To check the behavior and stability of the DMPC bilayer and the peptide alone in water, simulations with the same setup as that in the multicomponent simulations were performed for each isolated component. Furthermore, the effect on the insertion mechanism of the lipid tails was assessed by simulating the peptide without lipid tails under the same conditions as those for the lipid-modified peptide.

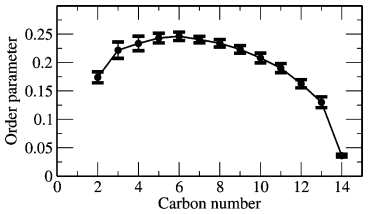


Figure 3. Deuterium order parameters of hydrocarbon tails from the 6 ns DMPC trajectory without peptides. The error bars are calculated by dividing the last 4 ns of the trajectory in segments of 1 ns.

Peptide in Water. One factor that may affect the rate of peptide insertion is its structural properties in water. To assess the behavior of the peptide in water, a short simulation of 2 ns was performed in a cubic box of TIP3 waters. As may be expected for a system with no charged or polar side chains, the peptide quickly adopted a collapsed conformation. It buried a large part of its hydrophobic surface by "sequestering" the apolar amino acids, Met and Leu, with the lipid hydrocarbon chains (data not shown). The short simulation does not allow one to describe the accessible conformational space.

DMPC in Water. A 6 ns trajectory of the DMPC bilayer was run to investigate the bilayer behavior in the absence of the peptides and to facilitate structural comparison with simulations in the presence of the ras peptide. The trajectory was stable, and the bilayer structural properties were generally the same as in previous reports.[55] The bilayer thickness (D_{PP}) is a useful parameter for estimating bilayer structural changes upon protein insertion. The average D_{PP} value calculated from the simulation without peptides (37.0 Å, see Table 4) is in good agreement with previous calculations using the same protocol (35.9 Å)[55] and is close to the experimental thickness (i.e., the average distance between lipid headgroups at each monolayer measured by electron density profile method) of 36.0 Å.[61] The deuterium order parameters (S_{CD}, eq 2), calculated from the simulation (Figure 3), are typical of other DMPC simulations[62] and in very good agreement with experimentally measured values at 30 °C.[52] Possible structural alterations after peptide insertions can therefore be safely attributed to changes arising from the insertion of the peptide.

Peptide without Lipid Tails. Three trajectories with the same starting conditions as those in trajectories A_2, A_3, and A_{1a}, but lacking the palmitoyl and hexadecyl groups, were run for 2.5 ns each. Four of the six peptides fully dissociated into water. For the remaining two peptides, most of the backbone and side chains moved toward bulk water within 2.5 ns, with only the N-terminus maintaining contact with the headgroup region of the bilayer. This suggests that lipid modifications are essential for membrane insertion.

Peptide Insertion. The location of the geometric center of the peptides and the initial number of peptide−DMPC contacts were used to monitor the insertion process and peptide−bilayer interactions. Table 2 summarizes the initial and final peptide locations, as well as the initial number of ras acyl carbon atoms in contact with those of DMPC. Initial peptide positions varied between 41 (completely in water) and 16 Å (partly inserted).

(55) Petrache, H. I.; Grossfield, A.; MacKenzie, K. R.; Engelman, D. M.; Woolf, T. B. *J. Mol. Biol.* **2000**, *302*, 727−746.
(56) Hoover, W. G. *Phys. Rev. A* **1985**, *31*, 1695−1697.
(57) Patra, M.; Karttunen, M.; Hyvonen, M. T.; Falck, E.; Lindqvist, P.; Vattulainen, I. *Biophys. J.* **2003**, *84*, 3636−3645.
(58) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089−10092.
(59) Petrache, H. I.; Tu, K.; Nägle, J. F. *Biophys. J.* **1999**, *76*, 2479−2487.
(60) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996**, *14*, 33−38.
(61) Nagle, J. F.; Tristram-Nagle, S. *Biochim. Biophys. Acta* **2000**, *1469*, 159−195.
(62) Moore, P. B.; Lopez, C. F.; Klein, M. L. *Biophys. J.* **2001**, *81*, 2484−2494.

Table 2. Initial and Final Peptide Positions[a]

		A_0	A_1	A_2	A_3	A_{1a}	C_0	C_1	C_2	C_3	C_{1a}
Number of Ras Acyl Carbon Atoms in Contact with DMPC Acyl Carbons at the Start[b]											
P_1	R_1	0	**7**	**15**	**15**	**15**	0	1	**4**	**8**	**7**
	R_2	0	3	**10**	**12**	**0**	0	2	**7**	**11**	0
P_2	R_1	0	0	**0**	**6**	**0**	0	**1**	**8**	**11**	**15**
	R_2	0	4	**11**	**15**	**7**	0	**11**	**11**	**16**	**16**
Distances between Peptide Geometric Centers and the Center of Bilayer (initial → final) (Å)[c]											
P_1		39 → 39	26 → 22	**19 → 10**	**17 → 10**	**22 → 13**	41 → 44	30 → 38	**24 → 9**	**21 → 9**	24 → 23
P_2		34 → 34	28 → 33	**23 → 8**	**19 → 12**	**23 → 9**	41 → 52	**22 → 15**	**19 → 10**	**16 → 7**	**15 → 11**

[a] P_1 and P_2 represent peptides 1 and 2, respectively, whereas R_1 and R_2 represent the palmitoyl and hexadecyl lipid tails of the ras peptide, respectively. [b] Contact is present if the ras acyl carbon is within 5 Å of at least one DMPC acyl carbon. The maximum possible number is 16 (i.e., the lengths of the palmitoyl and hexadecyl chains). [c] Initial and final peptide positions are defined by the location of the peptide geometric center in the first and last snapshots of the simulations [absolute values, as measured from the bilayer center ($z = 0$), are indicated]. The initial contacts that led to insertion (i.e., to the maximum possible number) and the corresponding locations are in bold.

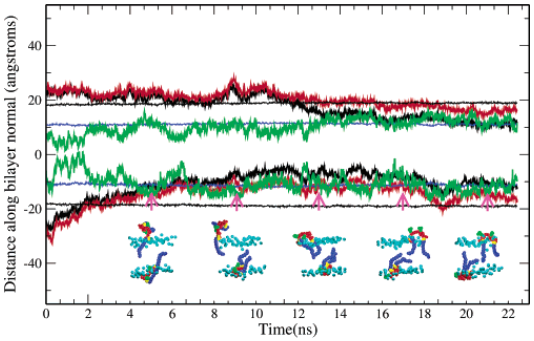


Figure 4. Time series and snapshots show the movement of the peptide toward the bilayer−water interface followed by insertion in trajectory A_{1a}. The cylindrical potential to prevent lateral motion of the peptide (see Methods) was removed at 6.5 ns. The large apparent displacement of P_1 between the snapshots at 13 and 17 ns is an effect of the periodic boundary conditions. Lines above and below $z = 0$ (the bilayer center) are for peptides 1 and 2, respectively. Thick lines: red, backbone geometric center; black, all atom geometric center; green, peptide acyl chain length. Thin lines: blue, DMPC acyl chain length; black, average phosphorus atom locations. The insets show the peptide insertion levels at selected time points during the simulations. Color codes are as in Figure 1B.

Whenever the peptide was placed near the interface with an average distance from the bilayer center of <25 Å (~7 Å from the average phosphorus atoms location), a subsequent insertion was observed (shown in bold, Table 2).

Figure 4 shows the progress of peptide insertion during the longest trajectory (A_{1a}). The peptides slowly approach the bilayer (at times even immersing in the hydrophobic region; heavy black lines) and eventually stabilize, with the backbone occupying the interfacial region. In agreement with Fourier transform infrared experiments,[14] the acyl chain lengths of the ras peptides equilibrate at lengths equivalent to those of the DMPC lipids tails, but the length fluctuations are much larger in the former. As can be seen from the insets of Figure 4, the peptide−DMPC hydrophobic contacts progressively increase along the trajectories. The two peptides move independently of each other; initially, they occupied similar locations in the middle of each monolayer but then translated to opposite corners of the monolayers' cross sectional area (inset).

Furthermore, the calculated center of mass of the peptides and the monolayers in the unconstrained simulations indicate that while the motion of the two monolayers relative to each other is negligible, there is a substantial lateral movement of each peptide independently of the other (data not shown). All of the simulations where insertions have been achieved show similar behavior. These observations indicate that once the peptides are inserted, the previous history (including the starting position and orientation) is forgotten. The mobility of the peptide at the membrane surface is qualitatively similar to that seen in a recent MD study of a membrane-anchored single-lipidated peptide.[63,64] Furthermore, the values of the two-dimensional diffusion constant on the membrane plane computed from the trajectories,[35] $\sim 10-30 \times 10^{-8}$ cm^2 s^{-1} for the DMPC and $\sim 6-15 \times 10^{-8}$ cm^2 s^{-1} for the peptide, suggest that the peptide moves slightly slower than individual lipids. The diffusion of the DMPC lipids is within the range of excimer experiments.[65]

Comparison of the trajectories, based on the contact criteria, may give a more-detailed picture of the insertion process. Although insertion was not observed in the simulations where no initial peptide−DMPC contacts were present (A_0 and C_0), a minimum of roughly 5−7 inserted carbon atoms per acyl chain is sufficient to spontaneously lead to full insertion. A smaller number of ras acyl carbons contacting DMPC acyl carbons did not result in insertion within the time scale of the simulations. For example, while seven carbon−carbon contacts between ras and the DMPC lipids resulted in insertion (e.g., P_2 in A_{1a} and in C_{1a}), three or four contacts could not stabilize the ras peptide in the bilayer and led to desorption to water (P_2 in A_1 and P_1 in C_1). Interestingly, even a chain with no or few initial contacts subsequently inserts if the other chain is involved in a sufficient number of initial contacts with the bilayer (e.g., P_2 in A_2 and in C_1). As an example in trajectory A_2, the palmitic chain (R_1) of P_2 was in water at the start and early stages of the simulation; its insertion began at about 9.3 ns, and after fluctuating for about 2 ns, it made a stable association from 11 ns onwards (Figure 5). In total, five cases of single chain preinsertion were used to investigate the behavior of the complex where only a single chain of a peptide was initially in contact with the bilayer [A_{1a}, $R_1(P_2)$ and $R_2(P_1)$; A_2, $R_1(P_2)$; C_1, $R_1(P_2)$; C_{1a}, $R_2(P_1)$]. It was anticipated that the palmitoyl chain R_1 (in trajectories A_{1a}, A_2, and C_1) and the hexadecyl chain R_2 (in trajectories A_{1a} and C_{1a}) would insert spontaneously. A total of three and one spontaneous insertions of the palmitoyl and hexadecyl chains were observed, respectively, whereas the hexadecyl group did not insert in the 20 ns simulation of C_{1a} (Table 3). Despite the limited statistics, it appears that there is no significant difference in the times

(63) Nagle, J. F. *Biophys. J.* **1993**, *64*, 1476−1481.
(64) Jensen, M. O.; Mouritsen, O. G.; Peters, G. H. *Biophys. J.* **2004**, *86*, 3556−3575.
(65) Blume, A. Dynamic Properties. In *Phospholipid Handbook*; Cevc, G., Ed.; Marcel Dekker: New York, 1993; pp 455−552.

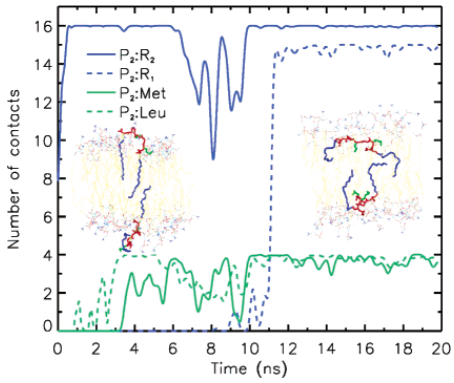


Figure 5. Peptide−DMPC hydrophobic interactions along trajectory A_2. Interactions are defined as the number of ras methylene/methyl carbons and Leu and Met side chain heavy atoms, within 5 Å of DMPC methylene/methyl carbons. For clarity, the data shown are only for peptide 2. The insets represent the system before (left, at 2 ns) and after (right, at 20 ns) complete insertion of peptide 2 (shown in the lower leaflet of the insets). The hydrophobic groups of ras interacting with the DMPC hydrophobic core are colored in blue and green; the DMPC lipid tails are in yellow, and the headgroups are in "standard" atom colors.

Table 3. Approximate Insertion Times (in nanoseconds) of Palmitoyl and Hexadecyl Groups of the Peptide[a]

chain (peptide)	time	trajectory
$R_1(P_2)$	2.4 (0.0)	C_1
$R_1(P_2)$	11.3 (9.3)	A_2
$R_1(P_2)$	7.0 (1.5)	A_{1a}
$R_2(P_1)$	12.3 (11.9)	A_{1a}

[a] Insertion times are measured for the insertion (distance of <5 Å from any DMPC acyl carbon) of at least 8 of the 16 ras acyl carbons. Time points where initial fluctuating contacts were made are shown in parentheses.

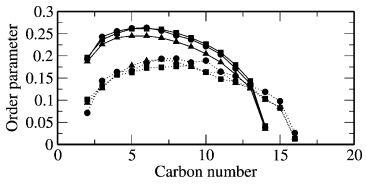


Figure 6. Deuterium order parameters of hydrocarbon tails for DMPC (solid lines) and peptide lipids (dotted lines) after insertion of the ras peptide. Circles, boxes, and triangles are for simulations C_2, A_2, and A_3, respectively.

needed for the Cys181 palmitoyl or Cys186 hexadecyl groups' spontaneous insertion. Moreover, as mentioned above, a limited preinsertion of one or both of the chains is sufficient to trigger insertion and stabilization of the whole peptide.

Lipid Bilayer Structure. Lower order parameters were observed for the 16:0 ras chains compared with the 14:0 DMPC chains.[14] The order parameters from the last 10 ns of some of the simulations that resulted in peptide insertion are shown in Figure 6. It is clear from the figure that the peptide lipid chains are significantly less ordered compared with the DMPC chains, which is in very good agreement with experiments.[14] The difference in the calculated average order parameters between the DMPC chains and the ras chains is 0.060, compared with the experimental value of 0.057. The effective length of saturated hydrocarbon chains is proportional to the average order parameters.[63] The results show that the longer 16:0 ras chains decrease their order upon membrane insertion to match the length of the shorter 14:0 DMPC lipids. In fact, the lengths of the ras and DMPC chains become close to each other after the stabilization of the peptide in the host lipids (Figure 4).

Table 4 compares the D_{PP} values calculated for the free DMPC simulation with those of the "insertion" simulations. In all of the insertion simulations, the global thickness of the bilayer has increased by about 1 Å: an average of 38.1 Å in the presence of the peptide, compared with 37 Å for a pure DMPC simulation. This is consistent with several ^{2}H NMR and ESR experiments which showed that hydrophobic peptides increase bilayer thickness due to hydrophobic mismatch.[66] Observed globally, therefore, the structure of the bilayer is slightly perturbed by the insertion of the peptide. However, the average chain lengths (L_C, eq 1) in simulations with and without peptides are similar (Table 4).

Besides the average thickness, the local effect of the peptide on the lipid bilayer is of special interest. The average distance, D_P, of the P atoms from the bilayer center (at $z = 0$) was calculated for those P atoms that are close to the peptide (a cutoff of 8 Å) and compared with the average distance for the rest of the P atoms. In all of the simulations with insertions, the D_P in the vicinity of the peptide is decreased with respect to the average value (Table 4). In contrast, the rest of the bilayer responds by increasing its D_P. Petrache et al. observed that the bilayer thickness around the monomeric form of glycophorine A decreases, as in the ras peptide (Table 4), while an increase was observed in the dimer.[55] The change of the bilayer structure due to the insertion of peptides is a complex process involving both the length and the tilt of the peptide and the lipid.[67] The above observation therefore requires a further investigation beyond the scope of this paper.

Peptide Structure. Structural analysis of the peptide backbone, in terms of its fluctuations, presence of hydrogen bonds, and other observables, revealed that the peptide does not assume a regular secondary structure before or after membrane insertion. No backbone hydrogen bonds were observed in the simulations. The root-mean-square fluctuations (RMSF) of the C_α atoms after complete insertion (0.5−1 Å for the last 2 ns of the simulations) suggest that the peptide is rather rigid (see the thickness of the tubes in Figure 7). The peptide backbone is extended, and several conformations are observed. Furthermore, while the overall backbone and side chain membrane localization of the peptide is similar among the trajectories, cluster and principal component analyses indicate that each trajectory samples a rather different region of conformational space with few overlaps (data not shown). Comparison of the peptide structures averaged over the last 2 ns in terms of root-mean-square deviations (RMSD) also shows the same behavior (Table 5). Although the simulation times may be insufficient to allow conformational interconversion, these results indicate that a unique peptide structure may not be required for bilayer adsorption. The lack of a well-defined three-dimensional structure may be biologically relevant since it increases the likelihood of productive encounters between the peptide/protein and plasma membrane.[69,70] However, the orientation of each side chain is important for the binding, and the backbone conformation should accommodate this requirement. Figure 7 (bottom left) shows the conformational change

(66) de Planque, M. R.; Greathouse, D. V.; Koeppe, R. E., II.; Schafer, H.; Marsh, D.; Killian, J. A. *Biochemistry* **1998**, *37*, 9333−9345.
(67) Petrache, H. I.; Killian, J. A.; Koeppe, R. E.; Woolf, T. B. *Biophys. J.* **2000**, *78*, 324A.

Table 4. Structural Properties of the DMPC Bilayer[a]

	without peptides	A_2	A_3	C_2	C_3
D_{PP}	37.0 ± 0.4	38.0 ± 0.4	38.1 ± 0.3	38.1 ± 0.3	38.3 ± 0.3
L_C	11.3 ± 0.2	11.4 ± 0.2	11.2 ± 0.2	11.5 ± 0.2	11.6 ± 0.2
D_P (bound, leaflet 1)		18.2 ± 0.8	18.5 ± 0.7	17.6 ± 1.1	17.8 ± 0.8
D_P (rest, leaflet 1)		19.3 ± 0.3	19.8 ± 0.3	19.6 ± 0.2	19.4 ± 0.3
D_P (bound, leaflet 2)		17.5 ± 0.8	18.3 ± 1.0	17.9 ± 0.8	16.7 ± 0.7
D_P (rest, leaflet 2)		19.4 ± 0.3	18.6 ± 0.3	18.9 ± 0.3	19.8 ± 0.3

[a] Data are averages over the last 4 ns of each trajectory (±, standard deviations). The bilayer thickness (D_{PP}) is the distance of the average P atom locations at each monolayer. D_P is the distance of the average P atom location of a monolayer from the bilayer center, calculated for all of the P atoms, closer by at least 8 Å to any peptide heavy atom (bound), and for the remaining P atoms (rest). The D_P values are given for leaflets 1 and 2 since the level of peptide insertion at each leaflet may differ. All values are given in angstroms.

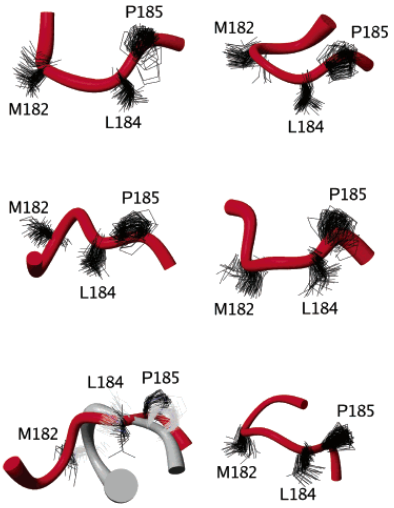


Figure 7. Average backbone structures of the peptides during the last 10 ns of trajectories A_{1a} (top), A_2 (middle), and A_3 (bottom); peptide 1 is shown in the left-hand-side and peptide 2 in the right-hand-side. The thickness of the tubes represents the root-mean-square fluctuations of the backbone. Also shown are the side chains: from left to right, Met182, Leu184, and Pro185. The bottom left structure shows the average structures before (gray) and after (red for the backbone and black for side chains) the conformational transition (see text for details). Figure made with MOLMOL.[68]

Table 5. C_α RMSD Values of the Peptide Structures in Different Insertion Simulations (angstroms)[a]

	A_2:P_1	A_2:P_2	A_3:P_1	A_3:P_2	A_{1a}:P_1	A_{1a}:P_2
A_2:P_1		2.6	1.9	3.4	2.3	3.3
A_2:P_2			3.2	2.3	1.4	2.4
A_3:P_1				4.0	2.9	4.4
A_3:P_2					2.7	1.7
A_{1a}:P_1						2.5

[a] Data for the simulations with an acetylated N-terminus are shown. The C_α RMSD values were calculated for structures averaged over the last 2 ns after complete peptide insertion.

of the backbone accompanying the rotation of the Leu side chain from water-exposed to the interior of the bilayer. Similar structural transitions are observed during the rotation and insertion of the Met side chain (data not shown). Experimentally, the amide bands in a Fourier transform infrared spectrum of the peptides at two polarizations were different, which indicates that the peptides in the sample adopt a nonrandom orientation.[14]

(68) Koradi, R.; Billeter, M.; Wuthrich, K. *J. Mol. Graphics* **1996**, *14*, 29−32, 51−55.
(69) Wright, P.; Dyson, H. *J. Mol. Biol.* **1999**, *293*, 321−331.
(70) Caflisch, A. *Trends Biotechnol.* **2003**, *21*, 423−425.

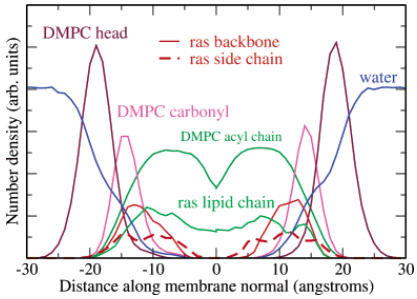


Figure 8. Number density averaged over the last 2 ns of all of the simulations that resulted in peptide insertion. Because of the different levels of peptide insertion in the earlier stages of the simulations, the average was taken in the trajectories with complete insertion and only for the last 2 ns.

Peptide Environment Interactions. Differences in the neutron scattering length density profiles between DMPC, with and without the ras peptide, were used to determine the insertion depth and distribution of the ras lipid chains as well as the distribution of backbone and side chains.[14] The ras lipid chains were found to be located in the middle of the membrane, approximately distributed within 10 Å of the bilayer center. The backbone and side chains reside in the headgroup/glycerol/upper chain region. The average number densities of the peptide lipid chains, side chains, backbone, and water, calculated from our simulations (Figure 8), strongly support these experimental findings. While the peptide lipid chains populate the interior of the bilayer and the side chains populate the upper chain region (peak at 11 Å), the backbone resides in the membrane−water interface (with a maximum at ∼13 Å) close to the DMPC carbonyl oxygens. The side chains show large variations populating the region beneath the DMPC carbonyl oxygens and the upper hydrophobic region of the bilayer. The distribution of the side chains populating the hydrocarbon region is similar to that obtained by Tieleman and colleagues for a Trp residue in their MD study of Arg/Lys containing interfacial pentapeptides partitioned in a solvated DOPC bilayer.[37] In contrast, they found that the charged Arg/Lys (as well as a Leu side chain) populates the interfacial region.

The distributions are governed by specific atomic interactions of the peptide with its environment. The carbon atoms of the ras lipid chains are in contact with carbons of the DMPC host matrix (Figures 4 and 5), contributing to stable association. In addition to the ras lipid chains, the apolar residues, Met and Leu, are involved in van der Waals interactions with the DMPC lipid chains (Figure 5). From the radial distribution functions (data not shown), polar interactions involving the amide and

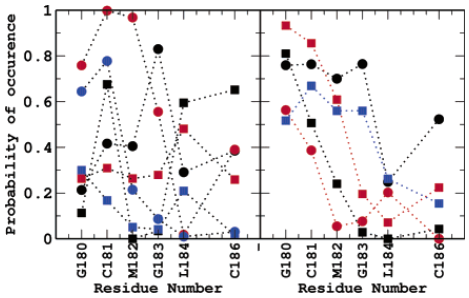


Figure 9. Probability of the occurrence of backbone amide nitrogen−phosphate oxygen hydrogen bonds in the last 10 ns of the simulations. A distance cutoff of 3.5 Å between N and O atoms was used. Left panel represents simulations with the acetylated N-terminus (A_2 in red, A_3 in black, and A_{1a} in blue). The right panel is for simulations with the charged N-terminus (C_2 in red, C_3 in black, and C_{1a} in blue). Circles are for peptide 1 and squares for peptide 2. The dotted lines are drawn only for clarity.

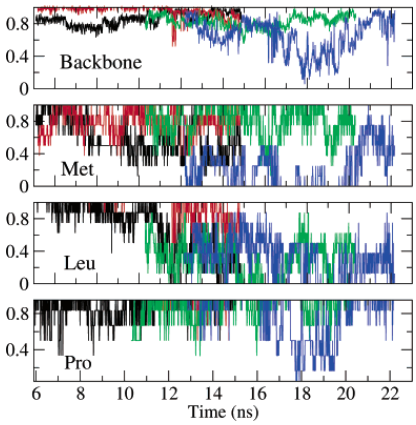


Figure 10. Number of ras atoms in contact with water during peptide insertion normalized over the number of contacts in a fully solvated extended conformation. A contact is defined as the number of ras heavy atoms within 5 Å of water oxygen atoms. The last 10 ns of A_3, C_3, A_2, and C_2 are shown in black, red, green, and blue, respectively.

carbonyl groups of the peptide backbone and the phosphate and choline groups of DMPC, respectively, anchor the peptide at the lipid−water interface. Hydrogen-bond interactions, particularly by the N-terminal side of the peptide (Figure 9), play a significant role. The plots in Figure 10 indicate the interactions of the different parts of the peptide with water in the last 10 ns of the trajectories. As expected, the Met and Leu side chains lose significant contact with water as they progressively penetrate the lipid membrane. The backbone and Pro side chain, however, remain in constant interaction with solvent. The interactions of the ras peptide with the DMPC bilayer and water are finely balanced. However, it is clear from the number of ras and DMPC acyl carbon contacts that nonpolar interactions provide the major driving force for association. The backbone and the Pro ring help maintain the location of the peptide in the interfacial region.

Peptide Stabilization in the Bilayer. Insertion of singly myristoylated or palmitoylated peptides containing positive charge clusters into bilayers with negatively charged interfaces is assisted by attractive electrostatic interactions. However, complete insertion of their lipid tails into the hydrophobic region is prevented by unfavorable electrostatic desolvation penalty.[71,72] In contrast, the completely hydrophobic polypeptide from the C-terminus of the N-ras protein was shown previously,[14] and here, to insert deep into the DMPC bilayer. This leads to a larger gain in hydrophobic stabilization compared with charged peptides. Although the insertion of the ras lipid chain is driven by hydrophobic interactions (the polar backbone−headgroup interactions playing a passive role in the initial contact formation; see the section on peptide insertion), once inserted, the backbone of the peptide contributes to the stability of the complex by interacting with the DMPC headgroups (Figure 9). Hydrogen-bonding interactions between the backbone amides (mainly involving the N-terminal segment, particularly the palmitoylated Cys181) and the phosphate oxygen atoms are frequent. The hydrogen-bonding potential appears to be slightly enhanced in the charged form of the peptide. In general, however, there was no significant difference between the charged and acetylated forms of the peptide in their interaction with the bilayer, suggesting that the hydrophobic interactions are the most crucial. This is further supported by the observed interactions between the lipid tails of the ras peptide and the DMPC. Note that the transfer from aqueous solution into a nonpolar environment of each CH2 group of lipid-modified proteins contributes about 0.8 kcal mol^{-1} of hydrophobic stabilization,[45,73] suggesting a large energetic contribution from the lipid tail interactions. Further hydrophobic stabilizations due to interactions between the apolar side chains and the DMPC hydrocarbon tails ensure a strong association.

As seen in the previous section, Met and Leu side chains orient toward the hydrocarbon core, while Pro and the backbone preferentially populate the membrane−water interface. This orientational scenario is in agreement with the experiment-based hydrophobicity scales of Wimley and White (WW scales) obtained from the partitioning of small peptides into lipid vesicles[74] and octanol.[75] The former can be interpreted as the free energy of transfer of whole peptides from water to membrane interfaces ($\Delta G_{w\rightarrow if}$) and the latter from water to the hydrocarbon (HC) core ($\Delta G_{w\rightarrow oct}$). The difference gives the relative free energy of transfer from interface to HC ($\Delta G_{if\rightarrow oct}$). When the lipid modifications are excluded, the total $\Delta G_{w\rightarrow if}$ is −0.8 kcal mol^{-1}; so, the peptide would slightly favor binding to the interface. The value of $\Delta G_{w\rightarrow oct}$ is 0.48 kcal mol^{-1}, and consequently, the free energy of interface-to-HC transfer ($\Delta G_{if\rightarrow oct}$) is 1.28 kcal mol^{-1}, such that the insertion of the nonlipid-modified peptide into the HC region would be unfavorable. The backbone is the major source of the unfavorable peptide−HC interaction as its contribution ($\Delta G_{if\rightarrow oct}^{backbone}$) is 5.6 kcal mol^{-1} (note that $\Delta G_{w\rightarrow if}^{backbone}$ and $\Delta G_{w\rightarrow oct}^{backbone}$ are 1.2 and 2.0 kcal mol^{-1} per residue, respectively[76]). However, insertion into the HC can be facilitated by side chain reorientations. When $\Delta G_{if\rightarrow oct}$ of the side chains is derived from the WW scales, Met and Leu, with −1.24 and −1.49 kcal mol^{-1}, respectively, favor

(71) Buser, C. A.; Sigal, C. T.; Resh, M. D.; McLaughlin, S. *Biochemistry* **1994**, *33*, 13093−13101.
(72) Pool, C. T.; Thompson, T. E. *Biochemistry* **1998**, *37*, 10246−10255.
(73) Peitzsch, R. M.; McLaughlin, S. *Biochemistry* **1993**, *32*, 10436−10443.
(74) Wimley, W.; White, S. *Nat. Struct. Biol.* **1996**, *3*, 842−848.
(75) Wimley, W. C.; Creamer, T. P.; White, S. H. *Biochemistry* **1996**, *35*, 5109−5124.
(76) White, S. H. *FEBS Lett.* **2003**, *555*, 116−121.

the HC region, while the two glycines, with a $\Delta G_{if\rightarrow oct}$ value of 0.68 kcal mol^{-1}, favorably interact with the interface. A Pro side chain, with -1.11 kcal mol^{-1}, would be expected to transfer to the HC, but it remains in the interface due to the rigidity of its backbone. The two Cys side chains are expected to point into the HC with a free energy gain of -1.16 kcal mol^{-1}. Any lipid modification obviously increases this value. On the basis of the WW scales, therefore, the insertion into the HC region of the apolar amino acids, as well as the interfacial localization of the backbone (as observed from the simulations), is thermodynamically favored, as is the insertion of the ras lipid tails into the HC core.

Furthermore, the insertion and stabilization of the peptide in the phospholipid required only a limited number of ras acyl carbon atoms in contact with those of DMPC (between five and seven), and insertion of one chain leads to a fast spontaneous insertion of the other. While the former confirms the hydrophobic origin of the driving force of the association, the latter may relate to the fact that singly lipid-modified peptides associate with plasma membranes with shorter half-life times.[73,77]

Conclusions

The C-terminal Cys181−palmitoyl and Cys186−farnesyl chains play a fundamental role as membrane anchors of the human N-ras protein. Recently, a battery of spectroscopic techniques was used to investigate the membrane localization of a heptapeptide with the amino acid sequence of residues 180−186 containing the two lipid modifications.[14] Here, MD simulations were used for a detailed characterization of the late stages of the insertion mechanism. The position and orientation of the ras peptide and its components (backbone, side chains, and lipid chains) during the MD runs are consistent with the spectroscopic data.[14] The agreement between simulations and experiments allows the use of the former for an atomic level description of the peptide−membrane association. The simulation results highlight four aspects that go beyond the model obtained from experiments. First, a partial insertion of both, or even only one chain, is sufficient to trigger complete insertion and stabilization of the peptide in the membrane within the time scale of the MD simulations (10−20 ns). Second, the monolayer thickness is slightly smaller for phospholipids in contact with the inserted peptide and slightly larger far away from it. Because of the simplified system used in the simulations, it is not possible to speculate if this membrane deformation is relevant for signal transduction. Third, over the 10 ns time scale of the MD simulations, the peptide backbone is rather rigid and extended, but backbone conformations differing by up to 4.0 Å are observed in different MD trajectories. It is likely that multiple peptide conformations are energetically accessible for rapid binding of the human N-ras protein to the membrane. Finally, although it is difficult to speculate on the details of the insertion mechanism (because of the constrained MD used for the initial peptide positioning), a coarse-grained sequence of events consists of an initial contact between ras lipid tails and the DMPC acyl chains, followed by complete lipid insertion, and almost concomitant side chain reorientation and backbone reorganization.

After our paper was submitted, an MD study of a synthetic, cationic C_{14}-N-acylated peptide (myristoyl−HWAHPGGHHA−amide) inserted into a dipalmitoylphosphatidylcholine (DPPC) lipid bilayer was published.[64] The force field and simulated time scale are the same as in the present work, whereas the simulation protocol (constant surface tension and only one peptide per bilayer) and phospholipids (DPPC vs DMPC) are slightly different. Interestingly, the peptide mobility at the membrane surface, as well as the multiple backbone conformations and small structural fluctuations around them, is similar in the two studies, despite the differences in peptide sequence, length, and number of lipid tails.

Acknowledgment. We gratefully acknowledge Dr. R. Böckmann for useful discussions, critical reading of the manuscript, and help in preparation of some of the figures. We thank Dr. E. Paci, G. Interlandi, and F. Rao for interesting discussions, and Dr. G. Settanni for suggesting the lipidated ras peptide as an interesting system for simulations. The simulations were performed on the Matterhorn Beowulf cluster at the Computing Center of the University of Zurich. We thank C. Bollinger and Dr. A. Godknecht for setting up the cluster, and the Canton of Zurich for generous hardware support. This work was supported by the Swiss National Competence Center in Structural Biology (NCCR).

JA046607N

(77) Silvius, J. R.; l'Heureux, F. *Biochemistry* **1994**, *33*, 3014−3022.

10 Conclusion and outlook

A variety of computational techniques were employed in this thesis to investigate different biophysical issues, emphasizing that methods must be adapted to the problem under study. It was shown that computational approaches are powerful and versatile because they can describe systems at different levels of approximations. Although many arguments were treated, a large part of this work was devoted to the investigation of amyloid fibril formation. This argument, deeply involved into medical research, witnesses a tremendous interest of theoretical and computational research, and we believe that the coarse-grained model introduced above can address many further questions.

First, it is proposed to investigate the effects of molecular crowding on amyloid formation. It is not known whether intracellular deposition or extracellular accumulation of β-amyloid peptide promotes the pathological process [94]. The excluded volume induced by the molecular crowding of the cytoplasm influences the kinetics of fibrils polymerization [95, 96], but it is not clear whether and which intermediates are stabilized. Simulations of aggregation might be performed in the presence of spheres of different radii to approximate molecular crowding effects, and shed light on the influence of molecular crowding on kinetics and pathways of fibril formation.

Second, the interactions of amyloid aggregates with membrane will be studied. Many investigators have observed a lytic activity of amyloid oligomer on liposomes [97]. Formation of pore-like structures at the membrane surface seems a common feature of amyloid peptides [98]. For these reasons, the perturbation of cell membrane permeability induced by amyloid aggregates might be the origin for cytotoxicity [99]. Simulations of membrane and peptides can be helpful to depict which is the role of amyloid oligomers for membrane instability and eventually shed light on the identity of the toxic species.

Finally, the mechanisms of inhibition of amyloid aggregation is the ultimate goal of our research. Protecting the polypeptide from assembling into amyloid oligomers either by using a small molecule [100, 101] or a peptide [102, 103], has been revealed as a valid strategy against Alzheimer disease. Yet, the processes behind the fibril regression, and monomer solubilization are not understood, and molecular dynamics simulations of

the type discussed in this thesis might reveal under which conditions fibrils and their intermediates are depleted.

11 Acknowledgements

Here I want to thank the many people, relatives, friends and collegues who directly and indirectly contributed and supported this work. Il mio primo pensiero va a Nathalie, che mi è stata vicina durante questi cinque anni. Pur nelle difficoltà iniziali, causate dalla lontananza, la sua presenza è stata essenziale, se non vitale. Ha reso questo soggiorno a Zurigo un periodo indimenticabile. (Devo inoltre ricordare anche il suo apporto scientifico: l'idea del modello semplificato, mi è venuta grazie a lei!). Poi devo molto ai miei genitori, Velia e Guido, per il loro premuroso e caloroso sostegno, e ai miei fratelli Federico e Francesca, con i quali ho condiviso le decisioni salienti di questo periodo. Un grazie particolare anche a Cecile e ad Israel, sempre immancabilmente vicini. Thanks to my supervisor, Amedeo, who gave me the opportunity to work in his lab, and thanks to all the former and present member of the group and the collaborators, in particular: Andrea C., Enrico, Fabio P., Peter, Gianni, Michele, Alex, Marco C., Gian Gaetano, Emanuele, Gianluca, Francesco, Urs, Steffi, Fabian, Andrea P., Philipp, François, Beatrice, Marino, Pietro, Darek, Ting, Danzhi, Rainer, and Christiane. The support and the friendship of this team have been essential in many occasions. Finally I want to thank all the friends that I met in Zürich, which were so many that it is difficult to remember them all. They were simply fantastic, and they rendered special all my spare time. Here a short list: Marco M., Misu, Marcello, Fabio L., Claudio, Francesca, Fabio K., Luca R., Pier Luigi, Giorgia, Alessandro and Cinzia.

References

1. Zimmerman, S. B. & Minton, A. P. Macromolecular crowding: biochemical, biophysical, and physiological consequences. *Annu Rev Biophys Biomol Struct* **22**, 27–65 (1993).

2. Anfinsen, C. B. Principles that govern the folding of protein chains. *Science* **181**, 223–230 (1973).

3. Levinthal, C. Are there pathways for protein folding. *Journal de Chimique Physique* **65**, 44 (1968).

4. Fink, A. L. Natively unfolded proteins. *Curr Opin Struct Biol* **15**, 35–41 (2005).

5. Dunker, A. K. *et al.* Intrinsically disordered protein. *J Mol Graph Model* **19**, 26–59 (2001).

6. Tompa, P. Intrinsically unstructured proteins. *Trends Biochem Sci* **27**, 527–533 (2002).

7. Wright, P. E. & Dyson, H. J. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J Mol Biol* **293**, 321–331 (1999).

8. Selkoe, D. J. Folding proteins in fatal ways. *Nature* **426**, 900–904 (2003).

9. Engelman, D. M. *et al.* Membrane protein folding: beyond the two stage model. *FEBS Lett* **555**, 122–125 (2003).

10. White, S. H. Translocons, thermodynamics, and the folding of membrane proteins. *FEBS Lett* **555**, 116–121 (2003).

11. Hansen, J. & McDonald, I. *Theory of simple liquids* (Academic press, 1986).

12. van Gunsteren, W. F. & Berendsen, H. J. C. Computer simulation of molecular dynamics: Methodology, applications, and perspectives in chemistry. *Angewandte Chemie* **29**, 992–1023 (1990).

13. MacKerell, A. D. *et al.* All-atom empirical potential for molecular modeling and dynamics studies of protein. *Journal of Physical Chemistry B* **102**, 3586–3616 (1998).

14. Cornell, W. D. *et al.* A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **117**, 5179–5197 (1995).

15. Schuler, L. D., Daura, X. & van Gunsteren, W. F. An improved gromos96 force field for aliphatic hydrocarbons in the condensed phase. *Journal of Computational Chemistry* **22**, 1205–1218 (2001).

16. Jorgensen, W. L., Maxwell, D. S. & Tirado-Rives, J. Development and testing of the opls all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* **118**, 11225 –11236 (1996).

17. Jorgensen, W. L. & Tirado-Rives, J. Potential energy functions for atomic-level simulations of water and organic and biomolecular systems. *Proc Natl Acad Sci U S A* **102**, 6665–6670 (2005).

18. Roux, B. & Simonson, T. Implicit solvent models. *Biophys Chem* **78**, 1–20 (1999).

19. Jackson, J. D. *Classical Electrodynamics* (John Wiley and sons, 1962).

20. Nina, M., Im, W. & Roux, B. Optimized atomic radii for protein continuum electrostatics solvation forces. *Biophys Chem* **78**, 89–96 (1999).

21. W. C. Still, R. C. H., A. Tempczyk & Hendrickson, T. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* **112**, 6127–6129 (1990).

22. Wesson, L. & Eisenberg, D. Atomic solvation parameters applied to molecular dynamics of proteins in solution. *Protein Sci* **1**, 227–235 (1992).

23. Ferrara, P., Apostolakis, J. & Caflisch, A. Evaluation of a fast implicit solvent model for molecular dynamics simulations. *Proteins* **46**, 24–33 (2002).

24. Fraternali, F. & Gunsteren, W. F. V. An efficient mean solvation force model for use in molecular dynamics simulations of proteins in aqueous solution. *J Mol Biol* **256**, 939–948 (1996).

25. Lazaridis, T. & Karplus, M. Effective energy function for proteins in solution. *Proteins* **35**, 133–152 (1999).

26. Pierotti, R. A. A scaled particle theory of aqueous and nonaqueous solutions. *Chem. Rev.* **76**, 717–726 (1976).

27. Postma, J. P. M., Berendsen, H. J. C. & Haak, J. R. Thermodynamics of cavity formation in water. a molecular dynamics study. *Faraday Symp. Chem. Soc.* **17**, 55 (1982).

28. Sharp, K. A., Nicholls, A., Friedman, R. & Honig, B. Extracting hydrophobic free energies from experimental data: relationship to protein folding and theoretical models. *Biochemistry* **30**, 9686–9697 (1991).

29. Privalov, P. L. & Makhatadze, G. I. Contribution of hydration to protein folding thermodynamics. II. The entropy and Gibbs energy of hydration. *J Mol Biol* **232**, 660–679 (1993).

30. Nettels, D., Gopich, I. V., Hoffmann, A. & Schuler, B. Ultrafast dynamics of protein collapse from single-molecule photon statistics. *Proc Natl Acad Sci U S A* **104**, 2655–2660 (2007).

31. Müller, M., Katsov, K., & Schick, M. Biological and synthetic membranes: What can be learned from a coarse-grained description? *Physics Reports* **434**, 113–176 (2006).

32. Reynwar, B. J. *et al.* Aggregation and vesiculation of membrane proteins by curvature-mediated interactions. *Nature* **447**, 461–464 (2007).

33. Ryckaert, J., Ciccotti, G. & Berendsen, H. Numerical integration of the cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes. *Journal of Computational Physics* **23**, 327?341 (1977).

34. Caves, L. S., Evanseck, J. D. & Karplus, M. Locally accessible conformations of proteins: multiple molecular dynamics simulations of crambin. *Protein Sci* **7**, 649–666 (1998).

35. Pande, V. S. *et al.* Atomistic protein folding simulations on the submillisecond time scale using worldwide distributed computing. *Biopolymers* **68**, 91–109 (2003).

36. Sugita, Y. & Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters* **314**, 141–151 (1999).

37. Rao, F. & Caflisch, A. Replica exchange molecular dynamics simulations of reversible folding. *Journal of Chemical Physics* **119**, 4035–4042 (2003).

38. Cecchini, M., Rao, F., Seeber, M. & Caflisch, A. Replica exchange molecular dynamics simulations of amyloid peptide aggregation. *J Chem Phys* **121**, 10748–10756 (2004).

39. Sugita, Y. & Okamoto, Y. Molecular mechanism for stabilizing a short helical peptide studied by generalized-ensemble simulations with explicit solvent. *Biophys J* **88**, 3180–3190 (2005).

40. Sunde, M. & Blake, C. The structure of amyloid fibrils by electron microscopy and x-ray diffraction. *Adv Protein Chem* **50**, 123–159 (1997).

41. Rochet, J. C. & Lansbury, P. T. Amyloid fibrillogenesis: themes and variations. *Curr Opin Struct Biol* **10**, 60–68 (2000).

42. Lansbury, P. T. & Lashuel, H. A. A century-old debate on protein aggregation and neurodegeneration enters the clinic. *Nature* **443**, 774–779 (2006).

43. Chromy, B. A. *et al.* Self-assembly of abeta(1-42) into globular neurotoxins. *Biochemistry* **42**, 12749–12760 (2003).

44. Gong, Y. *et al.* Alzheimer's disease-affected brain: presence of oligomeric a beta ligands (addls) suggests a molecular basis for reversible memory loss. *Proc Natl Acad Sci U S A* **100**, 10417–10422 (2003).

45. Cleary, J. P. *et al.* Natural oligomers of the amyloid-beta protein specifically disrupt cognitive function. *Nat Neurosci* **8**, 79–84 (2005).

46. Rochet, J. C. & Lansbury, P. T. Amyloid fibrillogenesis: themes and variations. *Curr Opin Struct Biol* **10**, 60–68 (2000).

47. Kodali, R. & Wetzel, R. Polymorphism in the intermediates and products of amyloid assembly. *Curr Opin Struct Biol* **17**, 48–57 (2007).

48. Chiti, F. *et al.* Designing conditions for in vitro formation of amyloid protofilaments and fibrils. *Proc Natl Acad Sci U S A* **96**, 3590–3594 (1999).

49. Fändrich, M., Fletcher, M. A. & Dobson, C. M. Amyloid fibrils from muscle myoglobin. *Nature* **410**, 165–166 (2001).

50. Jimenez, J. L. *et al.* Cryo-electron microscopy structure of an sh3 amyloid fibril and model of the molecular packing. *EMBO J* **18**, 815–821 (1999).

51. Guijarro, J. I., Sunde, M., Jones, J. A., Campbell, I. D. & Dobson, C. M. Amyloid fibril formation by an sh3 domain. *Proc Natl Acad Sci U S A* **95**, 4224–4228 (1998).

52. Litvinovich, S. V. *et al.* Formation of amyloid-like fibrils by self-association of a partially unfolded fibronectin type iii module. *J Mol Biol* **280**, 245–258 (1998).

53. Fändrich, M. *et al.* Myoglobin forms amyloid fibrils by association of unfolded polypeptide segments. *Proc Natl Acad Sci U S A* **100**, 15463–15468 (2003).

54. Chiti, F. *et al.* Mutational analysis of the propensity for amyloid formation by a globular protein. *EMBO J* **19**, 1441–1449 (2000).

55. Dobson, C. M. Protein misfolding, evolution and disease. *Trends Biochem Sci* **24**, 329–332 (1999).

56. Christopeit, T. *et al.* Mutagenic analysis of the nucleation propensity of oxidized alzheimer's beta-amyloid peptide. *Protein Sci* **14**, 2125–2131 (2005).

57. O'Nuallain, B., Shivaprasad, S., Kheterpal, I. & Wetzel, R. Thermodynamics of a beta(1-40) amyloid fibril elongation. *Biochemistry* **44**, 12709–12718 (2005).

58. Tjernberg, L. O. *et al.* Arrest of beta-amyloid fibril formation by a pentapeptide ligand. *J Biol Chem* **271**, 8545–8548 (1996).

59. Ivanova, M. I., Sawaya, M. R., Gingery, M., Attinger, A. & Eisenberg, D. An amyloid-forming segment of beta2-microglobulin suggests a molecular model for the fibril. *Proc Natl Acad Sci U S A* **101**, 10584–10589 (2004).

60. Caflisch, A. Computational models for the prediction of polypeptide aggregation propensity. *Curr Opin Chem Biol* **10**, 437–444 (2006).

61. Chiti, F. *et al.* Kinetic partitioning of protein folding and aggregation. *Nat Struct Biol* **9**, 137–143 (2002).

62. Chiti, F., Stefani, M., Taddei, N., Ramponi, G. & Dobson, C. M. Rationalization of the effects of mutations on peptide and protein aggregation rates. *Nature* **424**, 805–808 (2003).

63. Gazit, E. A possible role for pi-stacking in the self-assembly of amyloid fibrils. *FASEB J* **16**, 77–83 (2002).

64. Kim, W. & Hecht, M. H. Generic hydrophobic residues are sufficient to promote aggregation of the alzheimer's abeta42 peptide. *Proc Natl Acad Sci U S A* **103**, 15824–15829 (2006).

65. Antzutkin, O. N. *et al.* Multiple quantum solid-state nmr indicates a parallel, not antiparallel, organization of beta-sheets in alzheimer's beta-amyloid fibrils. *Proc Natl Acad Sci U S A* **97**, 13045–13050 (2000).

66. Bond, J. P. *et al.* Assemblies of alzheimer's peptides a beta 25-35 and a beta 31-35: reverse-turn conformation and side-chain interactions revealed by x-ray diffraction. *J Struct Biol* **141**, 156–170 (2003).

67. Clark, P. L. Protein folding in the cell: reshaping the folding funnel. *Trends Biochem Sci* **29**, 527–534 (2004).

68. Conway, K. A., Harper, J. D. & Lansbury, P. T. Fibrils formed in vitro from alpha-synuclein and two mutant forms linked to Parkinson's disease are typical amyloid. *Biochemistry* **39**, 2552–2563 (2000).

69. Nilsberth, C. *et al.* The 'arctic' app mutation (e693g) causes alzheimer's disease by enhanced abeta protofibril formation. *Nat Neurosci* **4**, 887–893 (2001).

70. Cheng, I. H. *et al.* Aggressive amyloidosis in mice expressing human amyloid peptides with the arctic mutation. *Nat Med* **10**, 1190–1192 (2004).

71. Kheterpal, I. *et al.* Abeta protofibrils possess a stable core structure resistant to hydrogen exchange. *Biochemistry* **42**, 14092–14098 (2003).

72. Fändrich, M. *et al.* Apomyoglobin reveals a random-nucleation mechanism in amyloid protofibril formation. *Acta Histochem* **108**, 215–219 (2006).

73. Fändrich, M. Absolute correlation between lag time and growth rate in the spontaneous formation of several amyloid-like aggregates and fibrils. *J Mol Biol* **365**, 1266–1270 (2007).

74. Gosal, W. S. *et al.* Competing pathways determine fibril morphology in the self-assembly of beta2-microglobulin into amyloid. *J Mol Biol* **351**, 850–864 (2005).

75. Jansen, R., Dzwolak, W. & Winter, R. Amyloidogenic self-assembly of insulin aggregates probed by high resolution atomic force microscopy. *Biophys J* **88**, 1344–1353 (2005).

76. Southhall, N., Dill, K. & Haymet, A. A view of hydrophobic effect. *J. Phys. Chem. B* **106**, 521–533 (2002).

77. Huang, D. & Chandler, D. The hydrophobic affect and the influence of solute-solvent attractions. *J. Phys. Chem. B* **106**, 2047–2053 (2002).

78. Dolgikh, D. A. *et al.* Alpha-lactalbumin: compact state with fluctuating tertiary structure? *FEBS Lett* **136**, 311–315 (1981).

79. Ohgushi, M. & Wada, A. 'molten-globule state': a compact form of globular proteins with mobile side-chains. *FEBS Lett* **164**, 21–24 (1983).

80. Kuwajima, K. & Arai, M. *Mechanisms of protein folding*, chap. The molten globule state: the physical picture and biological significance, 138 (Oxford University Press, 2000).

81. Akasako, A., Haruki, M., Oobatake, M. & Kanaya, S. Conformational stabilities of escherichia coli rnase hi variants with a series of amino acid substitutions at a cavity within the hydrophobic core. *J Biol Chem* **272**, 18686–18693 (1997).

82. Binz, H. K. & Plückthun, A. Engineered proteins as specific binding reagents. *Curr Opin Biotechnol* **16**, 459–469 (2005).

83. Forrer, P., Binz, H. K., Stumpp, M. T. & Plückthun, A. Consensus design of repeat proteins. *Chembiochem* **5**, 183–189 (2004).

84. Bueno, M., Campos, L. A., Estrada, J. & Sancho, J. Energetics of aliphatic deletions in protein cores. *Protein Sci* **15**, 1858–1872 (2006).

85. Voigt, C. A., Gordon, D. B. & Mayo, S. L. Trading accuracy for speed: A quantitative comparison of search algorithms in protein sequence design. *J Mol Biol* **299**, 789–803 (2000).

86. Ventura, S. & Serrano, L. Designing proteins from the inside out. *Proteins* **56**, 1–10 (2004).

87. Bradley, P., Misura, K. M. S. & Baker, D. Toward high-resolution de novo structure prediction for small proteins. *Science* **309**, 1868–1871 (2005).

88. Desjarlais, J. R. & Handel, T. M. De novo design of the hydrophobic cores of proteins. *Protein Sci* **4**, 2006–2018 (1995).

89. Minetti, C. A. S. A. & Remeta, D. P. Energetics of membrane protein folding and stability. *Arch Biochem Biophys* **453**, 32–53 (2006).

90. Bechinger, B. Understanding peptide interactions with the lipid bilayer: a guide to membrane protein engineering. *Curr Opin Chem Biol* **4**, 639–644 (2000).

91. Ladokhin, A. S. & White, S. H. Interfacial folding and membrane insertion of a designed helical peptide. *Biochemistry* **43**, 5782–5791 (2004).

92. Scheffzek, K., Lautwein, A., Kabsch, W., Ahmadian, M. R. & Wittinghofer, A. Crystal structure of the gtpase-activating domain of human p120gap and implications for the interaction with ras. *Nature* **384**, 591–596 (1996).

93. Huster, D. *et al.* Membrane insertion of a lipidated ras peptide studied by ftir, solid-state nmr, and neutron diffraction spectroscopy. *J Am Chem Soc* **125**, 4070–4079 (2003).

94. Knobloch, M., Konietzko, U., Krebs, D. C. & Nitsch, R. M. Intracellular abeta and cognitive deficits precede beta-amyloid deposition in transgenic arcabeta mice. *Neurobiol Aging* (2006).

95. Shtilerman, M. D., Ding, T. T. & Lansbury, P. T. Molecular crowding accelerates fibrillization of alpha-synuclein: could an increase in the cytoplasmic protein concentration induce Parkinson's disease? *Biochemistry* **41**, 3855–3860 (2002).

96. Ellis, R. J. & Minton, A. P. Protein aggregation in crowded environments. *Biol Chem* **387**, 485–497 (2006).

97. Maltseva, E., Kerth, A., Blume, A., Mhwald, H. & Brezesinski, G. Adsorption of amyloid beta (1-40) peptide at phospholipid monolayers. *Chembiochem* **6**, 1817–1824 (2005).

98. Quist, A. *et al.* Amyloid ion channels: a common structural link for protein-misfolding disease. *Proc Natl Acad Sci U S A* **102**, 10427–10432 (2005).

99. Murphy, R. M. Kinetics of amyloid formation and membrane interaction with amyloidogenic proteins. *Biochim Biophys Acta* (2007).

100. Blanchard, B. J. *et al.* Efficient reversal of alzheimer's disease fibril formation and elimination of neurotoxicity by a small molecule. *Proc Natl Acad Sci U S A* **101**, 14326–14332 (2004).

101. Aisen, P. S. *et al.* A phase ii study targeting amyloid-beta with 3aps in mild-to-moderate alzheimer disease. *Neurology* **67**, 1757–1763 (2006).

102. Kokkoni, N., Stott, K., Amijee, H., Mason, J. M. & Doig, A. J. N-methylated peptide inhibitors of beta-amyloid aggregation and toxicity. optimization of the inhibitor structure. *Biochemistry* **45**, 9906–9918 (2006).

103. Yan, L.-M., Tatarek-Nossol, M., Velkova, A., Kazantzis, A. & Kapurniotu, A. Design of a mimic of nonamyloidogenic and bioactive human islet amyloid polypeptide (iapp) as nanomolar affinity inhibitor of iapp cytotoxic fibrillogenesis. *Proc Natl Acad Sci U S A* **103**, 2046–2051 (2006).